The Sound of Syntax: Finetuning and Comprehensive Evaluation of Language Models for Speech Pathology

Fagun Patel[1*], Duc Q. Nguyen[2*], Sang T. Truong[1*],
Jody Vaynshtok[3], Sanmi Koyejo[1], Nick Haber[1]
[1]Stanford University [2]National University of Singapore [3]Sound Speech and Hearing Clinic
∗ Co-first Authors

Abstract

According to the U.S. National Institutes of Health, more than 3.4 million children experience speech disorders that require clinical intervention. The number of speech-language pathologists (SLPs) is roughly 20 times fewer than the number of affected children, highlighting a significant gap in children's care and a pressing need for technological support that improves the productivity of SLPs. State-of-the-art multimodal language models (MLMs) show promise for supporting SLPs, but their use remains underexplored largely due to a limited understanding of their performance in high-stakes clinical settings. To address this gap, we collaborate with domain experts to develop a taxonomy of real-world use cases of MLMs in speech-language pathologies. Building on this taxonomy, we introduce the first comprehensive benchmark for evaluating MLM across five core use cases, each containing 1,000 manually annotated data points. This benchmark includes robustness and sensitivity tests under various settings, including background noise, speaker gender, and accent. Our evaluation of 15 state-of-the-art MLMs reveals that no single model consistently outperforms others across all tasks. Notably, we find systematic disparities, with models performing better on male speakers, and observe that chain-of-thought prompting can degrade performance on classification tasks with large label spaces and narrow decision boundaries. Furthermore, we study fine-tuning MLMs on domain-specific data, achieving improvements of over 30% compared to base models. These findings highlight both the potential and limitations of current MLMs for speech-language pathology applications, underscoring the need for further research and targeted development[1].

[1]To support continued progress, we publicly release our datasets, fine-tuned models, and benchmarking framework. Code: https://github.com/sangttruong/slp_benchmark, Dataset: https://huggingface.co/datasets/SAA-Lab/SLPHelmUltraSuitePlus.

1 Introduction

Speech and language pathologies (SLP) in children can significantly impact communication, academic development, and long-term social outcomes (Hitchcock et al., 2015; Foster et al., 2023). Early detection and intervention by speech-language pathologists are critical to mitigating these adverse effects (Gibbard et al., 2004; Centers for Disease Control and Prevention, 2024). Unfortunately, the availability of qualified clinicians is characterized by an uneven distribution across geographic and socioeconomic contexts, with only one expert for every 20 affected children, resulting in significant disparities in access to care and leading to "missing intervention" for many children who could benefit from timely support (U.S. National Institute on Deafness and Other Communication Disorders, 2025; Tucker and McKinnon, 2020). This gap highlights an urgent need for scalable and supportive technological solutions to augment clinicians' capabilities and expand the reach of vital interventions.

The shortage of qualified clinicians has led to significant gaps in diagnostic capacity, particularly in domains that require specialized expertise. Recent advancements in multimodal large language models (MLMs) present a promising opportunity to partially automate or augment diagnostic workflows (Lammert et al., 2025; Bhattacharya et al., 2024; Nagpal et al., 2025; Maqsood et al., 2024). Multimodal LLMs, such as GPT-4 and Gemini, exhibit state-of-the-art capabilities in speech processing and contextual reasoning. Effective integration of LLMs into clinical SLP workflow requires rigorous evaluation to establish their clinical validity (Cordella et al., 2025). This process relies on comprehensive and high-quality datasets that capture the variability of pediatric speech, especially disordered forms, and are annotated with clinical features. Currently, the evaluation is hindered by

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34896–34914
November 4-9, 2025

two key challenges: the scarcity of well-curated pediatric speech corpora and the lack of comprehensive evaluation frameworks for analyzing childrens speech (Suh et al., 2024).

In this study, we present a comprehensive evaluation of MLMs in SLP. We develop a systematic procedure for annotating symptoms and disorders in child speech, creating resources suitable for SLP-focused model evaluation and fine-tuning. The benchmark assesses models across five clinical scenarios covering a spectrum of tasks from foundational disorder detection to more granular symptoms, including Disorder Diagnosis, Transcription-Based Diagnosis, Transcription, Disorder Type Classification, and Symptom Classification. We systematically evaluate the capabilities and limitations of existing MLMs in SLP-relevant contexts and explore approaches for domain-specific adaptation. Our analysis reveals substantial performance gaps, with macro-F1 scores frequently falling below clinically acceptable thresholds, particularly on fine-grained tasks such as Disorder Type and Symptom Classification. To mitigate these shortcomings, we develop and assess fine-tuned MLMs, demonstrating significant improvements and advancing the state of the art on these specialized clinical tasks. Our contributions are summarized as follows.

- We release four curated pediatric speech datasets comprising approximately 30,000 speech samples across English and French, encompassing both typical and disordered speech. These datasets provide a publicly available, high-quality resource to support reproducible benchmarking in SLP.
- We introduce the first comprehensive evaluation framework for SLP, encompassing five essential clinical tasks. This framework enables consistent, task-aligned evaluation and facilitates direct comparison of speech LLM performance under a standardized protocol.
- We introduce fine-tuned speech LLMs that achieve state-of-the-art performance across all evaluated SLP tasks, illustrating the efficacy of domain-specific adaptation in enhancing diagnostic and transcriptional capabilities.
- We conduct extensive fine-grained analyses on model performance across various conditions, including demographic factors (gender, age), languages, and reasoning paradigms (e.g., Chain-of-Thought), offering deeper insights into model robustness and potential biases.

2 Related Works

Comprehensive AI Benchmarking Comprehensive benchmarking has been instrumental in advancing speech-health research. The ADReSS Challenge (Luz et al., 2020) established a balanced benchmark for Alzheimers detection from spontaneous speech, standardizing evaluation via F1 and MMSE-regression metrics. Similarly, the Childrens ASR Benchmark (Fan et al., 2024) introduced standardized splits and Whisper/Wav2Vec baselines for speech recognition in children aged 614, highlighting age-specific acoustic challenges. Nonetheless, systematic benchmarking of MLMs in clinical contexts remains limited.

Finetuning Audio Models Recent approaches have shown finetuning led to improvement in LLM audio understanding performance, especially for low-resource languages (Pillai et al., 2024). Models finetuned via approaches like Continued Pretraining (Ke et al., 2023) and Reasoning Preference Optimization (RPO) (Pang et al., 2024) have showcased improved reliability, factual accuracy, and cross-language generalization in clinical use cases (Kawakami et al., 2025). Additionally, fine-tuned LLMs have been shown to outperform classical DNN approaches across clinical use cases such as murmur detection (Florea et al., 2025). Furthermore, research has shown that while fine-tuning on specialized medical data enhances domain-specific knowledge, it can negatively impact a model's long-context understanding (Yang et al., 2024), highlighting the need for a balanced data composition during the fine-tuning process.

3 Method

3.1 Datasets, Annotation, and Core Scenarios

We use four datasets including Ultrasuite (Eshky et al., 2018), ENNI (Schneider et al., 2006), LeNormand (Le Normand, 1997), and Percept-GFTA (Benway et al., 2022). These datasets are publicly available, having been collected with informed consent and subsequently anonymized by their creators to protect participant privacy. The datasets encompass a range of child speech samples, both typical and disordered, and serve as the foundation for evaluating model performance

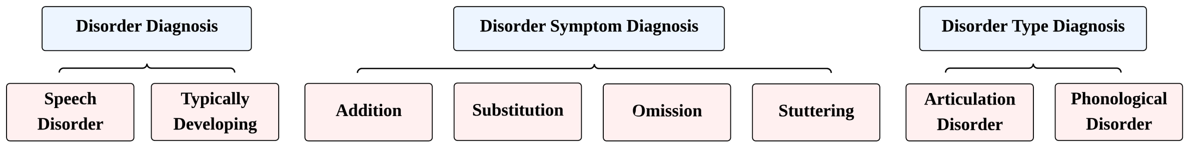

Figure 1: Taxonomy of classification tasks in SLPHelm. The benchmark includes three core diagnostic tasks: (i) disorder diagnosis, (ii) disorder symptom classification, and (iii) disorder type classification.

across diagnostic tasks. Their summary statistics are in Table 1. We randomly sample up to 1000 instances from each dataset for evaluation.

While foundational, existing datasets for children's speech-language pathology research are largely confined to transcription and binary classification of speech as either disordered or typical (Benway et al., 2022; Eshky et al., 2018; Le Normand, 1997; Schneider et al., 2006). This scope is critically insufficient for developing and validating models for more nuanced clinical applications, such as identifying specific disorder types or their underlying symptomscategories for which no large-scale, publicly available datasets currently exist. This lack of fine-grained annotation represents a significant bottleneck, impeding the development of automated tools that can support SLPs in differential diagnosis and therapy planning. To address this foundational gap, we collaborated closely with a certified SLP professional to develop a detailed annotation schema that captures both disorder types and their characteristic symptoms. For each speech sample, we assign the most prominent disorder type and symptom, prioritizing the most salient diagnostic features when multiple conditions may co-occur. Speech samples exhibiting no observable signs of speech disorder are annotated as typical. Our annotation protocol and chosen taxonomy are informed by clinical guidelines from the U.S. National Institutes of Health (Simon and Rosenbaum, 2016) and SLP best practices (American Speech-Language-Hearing Association, 2016). After initial manual labeling by our team, we conducted a verification phase in which the speech-language pathologist reviewed annotations to ensure consistency and clinical validity.

We evaluate five core tasks that collectively capture the essential stages of pediatric SLP, from initial screening to detailed diagnostic analysis: **(1) Disorder Diagnosis**, which assesses a model's ability to distinguish between typical and disordered speecha critical early triage step for prioritizing clinical resources; **(2) Transcript-based Diagnosis,** which serves as a baseline for diagnostic accuracy by testing the assumption that speech from children with disorders deviates from expected utterances. This approach operates by matching model-generated transcripts to required spoken text, which offers a minimal, interpretation-free method that could be readily deployed in clinical settings. By benchmarking against this heuristic, we quantify the value added by more sophisticated multimodal LLMs; **(3) Transcription**, which measures the fidelity of automatic speech recognition (ASR) systems on child with disordered speech, a prerequisite for downstream diagnostic and documentation tasks; **(4) Disorder Type Classification**, which probes whether models can differentiate between *articulation disorders*motor-based speech errors such as lispsand *phonological disorders*, which involve rule-based sound pattern errors such as omission of a part of a word shape (e.g., ca/cat) or consistent shifts in the place of production (e.g., /k/ → /t/). For example, a child who consistently replaces final consonants, such as saying "gape" for "gate," is exhibiting a pattern-based phonological process. A successful model must identify this underlying rule, distinguishing it from an articulation disorder where a child might struggle to physically produce the /t/ sound in any context; **(5) Disorder Symptom Classification**, a more granular task, requires models to pinpoint the specific clinical symptoms that constitute these error patterns. This includes identifying substitutions, such as the final consonant changing in the production of "gape" for "gate", or omissions, where a child might say "gore" instead of "gorge" by dropping the final sound. The task also includes identifying additions (the insertion of extra sounds) and stuttering (disruptions in speech fluency). Accurate identification of these individual symptoms is a critical prerequisite for downstream diagnostic analysis and informs the design of targeted therapy plans. Figure 1 illustrates an overview of classification tasks in our pipeline. Details prompts are presented in Appendix F.

Evaluation metrics for classification tasks include Macro F1, Micro F1, and Exact Match Accuracy, while transcription performance is assessed using Word Error Rate (WER), Match Error Rate (MER), and Word Information Preserved (WIP). The Macro F1 and Micro F1 scores were specifically chosen to provide a comprehensive assessment of classification performance. The Macro F1 score evaluates the model's average performance on each class equally, ensuring that performance on rare but clinically significant disorder categories is not overlooked. In contrast, the Micro F1 score aggregates performance across all individual samples, offering a measure of overall classification correctness.

We implement two distinct model inference pipelines. The first, referred to as the audio-to-LLM prompting pipeline, is designed for models with native multimodal capabilities (e.g., GPT-4o-Audio, Gemini 2.0 Flash). Here, raw audio inputs are passed directly to the model alongside a task-specific prompt. The second pipeline, termed transcription-based prompting, targets language-only models (denoted with the -transcribe suffix). Here, audio inputs are first transcribed using a base automatic speech recognition (ASR) model (e.g., Whisper or GPT-4's internal ASR), and the resulting text is embedded into a structured prompt for downstream reasoning. This two-pronged architecture compares models with native audio understanding to those using ASR-to-LLM pipelines, highlighting trade-offs between direct and transcription-mediated processing.

3.2 Models and Finetuning

We evaluate 15 speech models. Among the closed-source LLMs, we study GPT-4 family (4o-audio, 4o-mini-audio, 4o-transcribe, and 4o-mini-transcribe), Whisper, and the Gemini 2.0 family (2.0-flash, 2.0-flash-lite). For open-source models, we study Qwen families (2.5-omni-7b, 2.5-omni-3b, 2-audio-7b, audio-chat), the Phi-4, and IBM Granite series (3.3-8b, 3.3-3b, 3.2-8b). These models were chosen to cover a range of model sizes and families across both closed and open sources.

To investigate the impact of fine-tuning on model performance across multiple tasks, we explore two fine-tuning strategies. Our first strategy involves fine-tuning the model on a speech recognition task (Scenario 3, as described above), relying on the models intrinsic ability to transfer knowledge to improve performance on related tasks. In this setup, both typical and disordered speech samples are labeled with the same expected transcriptions. However, assigning identical transcriptions to acoustically distinct inputs may introduce ambiguity and limit the model's ability to learn disorder-specific patterns. To mitigate this, our second strategy - Finetuned w/ Markers - modifies the labeling of disordered speech by appending an asterisk to each word in its transcription. This lightweight labeling scheme serves to differentiate disordered speech from typical speech, thereby guiding the model to better recognize and transcribe disordered speech patterns without altering the overall task formulation. Details of fine-tuning prompts and hyperparameters are presented in Appendix D.

We hypothesize that fine-tuning on a general task (e.g., speech recognition) alone is insufficient to yield improvements on specialized clinical tasks unless the fine-tuning data contains explicit information relevant to those tasks. This stems from the theoretical premise that general-purpose models primarily optimize for surface-level acoustic-linguistic alignment, which may not encode the deeper, disorder-specific features, such as atypical phonological patterns or motor-based distortions, necessary for clinical inference (Shor et al., 2019; Dorfner et al., 2024).

4 Experiment

4.1 Per Scenario Results

Our findings suggest that both existing proprietary and open-source models currently fail to meet clinically acceptable performance thresholds. This limitation is likely attributable to the underrepresentation of disordered speech in training corpora, as such data is significantly less prevalent than typical speech samples available online. Existing FDA-approved diagnostic systems typically achieve F1 scores in the range of 0.80 to 0.85 (Fanni et al., 2023; Abràmoff et al., 2018), which serves as a practical standard for clinical viability. Furthermore, model performance varies across different task scenarios, highlighting the absence of a universally robust model that can consistently address the diverse requirements of pediatric SLP applications. Figure 2 presents an overview of the performance of all models.

Scenario 1: Disorder Diagnosis In the disorder diagnosis task, performance remains limited, with no model exceeding a micro F1 score of 0.56. The

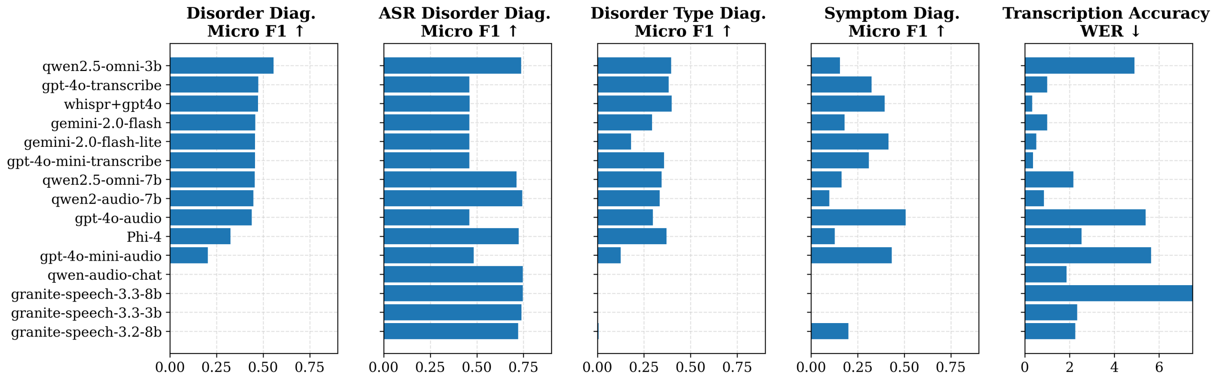

Figure 2: Metrics across all scenarios

best result is achieved by Qwen 2.5-Omni-3B, outperforming GPT-4o-Mini-Transcribe. The fact that these models use different pipelinesaudio-grounded vs. ASR+text suggests no clear advantage of one approach over the other. Smaller variants within each family exhibit similar performance. Audio-grounded Granite models perform poorly (F1 < 0.1), likely due to their pretraining focus on speech-to-text and translation tasks (IBM Granite Team, 2025).

Scenario 2: Transcription-based Diagnosis In this scenario, we evaluate a naive baseline that performs diagnosis on transcribed text rather than directly from audio. Counter-intuitively, this two-stage "transcribe-and-compare" approach does not degrade performance compared to end-to-end acoustic reasoning. In fact, for a majority of the models, it proves to be a superior strategy, yielding higher F1 scores. This surprising result suggests that potential error propagation from the ASR system is less of a hindrance than the inherent difficulty models face in performing complex diagnostic reasoning directly on acoustic features. Consequently, this simple transcription-based method, far from being a conservative lower bound, establishes itself as a highly competitive and often preferable baseline for diagnosing speech disorders.

Scenario 3: Transcription WER varies widelyfrom 0.33 to 6+, with the Gemini and OpenAI family of models performing the best. Importantly, transcription fidelity shows limited correlation with diagnostic accuracy: qwen2.5 omni 3b records a poor WER of 4.9 yet is consistently one of the strongest in classification tasks. These findings indicate that high-quality transcripts are neither necessary nor sufficient for dependable clinical reasoning.

Scenario 4: Disorder Type Classification Here, Whisper-GPT4o outperforms all models regardless of access or scale, hinting at architectural or pre-training advantages. The transcript-based inference consistently outperforms its audio-grounded counterparts. While for the larger GPT-4o model, audio-grounded inference performs close to its transcript-based counterpart, this trend reverses dramatically for the smaller 4o-mini model, where the ASR+LLM pipeline significantly outperforms the audio-grounded inference. This suggests that while large models can leverage the rich signals in raw audio, smaller models benefit substantially from the structured and potentially less noisy input of a transcript.

Scenario 5: Disorder Symptom Classification Accurate identification of these symptoms directly informs treatment goals and therapy design in speech-language pathology. GPT-4o leads in performance but still falls well short of clinically actionable accuracy. Moreover, transcription-based models underperform across all metrics, underscoring that critical acoustic cues needed for symptom detection are often lost or degraded during transcription.

4.2 Robustness Analysis

The Effect of Finetuning Fine-tuning large models can significantly enhance their performance on downstream tasks. Fine-tuning solely on automatic speech recognition (ASR) data, regardless of whether disordered speech is explicitly marked, leads to noticeable improvements in ASR-based tasks (Scenarios 2 and 3). However, not differentiating between typical and disordered speech introduces ambiguity in the input-label mapping, which degrades performance in other scenarios. Incorporating a simple asterisk mitigates this issue,

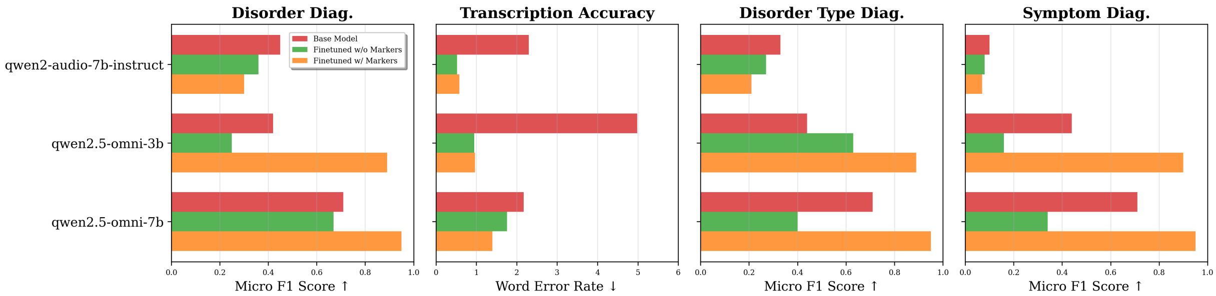

Figure 3: Model performance after finetuning

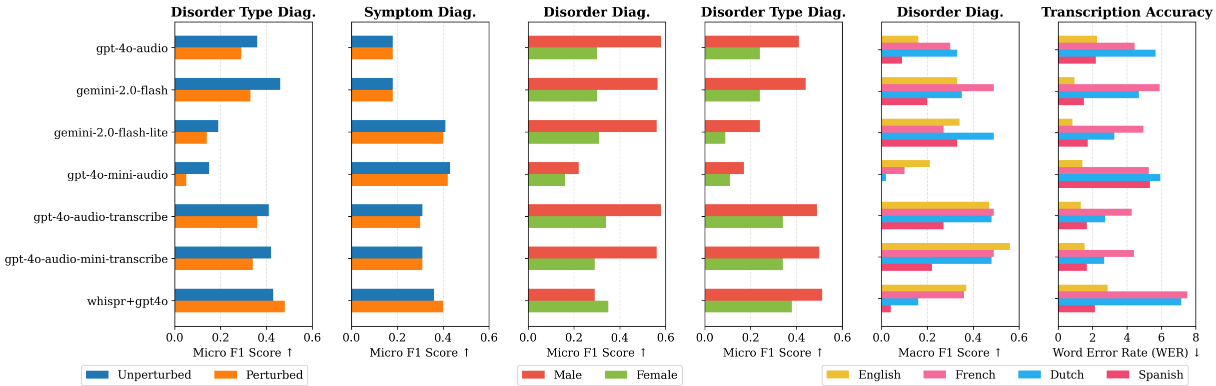

Figure 4: odel performance across robustness under noisy conditions, across gender, and languages

yielding more stable performance.

The Effect of Gender Figure 4 showcases that across two diagnostic tasks, the models exhibit a systematic gender performance gap that favors male speech. For both tasks, we run model evaluation on 1000 utterances for each gender on the UltraSuite dataset since it makes demographic identifiers available through its metadata. The results showcase a remarkably consistent pattern: almost every model posts a positive male-female differential. Notably, the gap is not confined to a particular modeling strategy; it appears in fully audio-grounded systems as well as in transcript-conditioned variants. Persistent speech-sound disorders are more commonly associated with boys than girls (Wren et al., 2016), which can lead to an imbalance in training data and downstream model performance. The magnitude of the divergence suggests practical consequences for clinical deployment, as female speech receives both lower sensitivity and lower precision across disorder categories. Taken together, the results underscore the need for targeted auditing and, potentially, gender-balanced fine-tuning to ensure equitable diagnostic performance across child speakers.

The Effect of Language Figure 4 shows a diverging pattern: classification accuracy shows no clear trend between languages, yet WER in French and Dutch is markedly worse. A plausible explanation lies in the way these systems were pre-trained. Their ASR components are heavily optimized on English text-to-speech pairs, so lexical recognition degrades when confronted with French/Dutch phonotactics, inflating WER. By contrast, the diagnostic classifiers operate on higher-level acoustic embeddings learned during large-scale audio pretraining that is largely language-agnostic (Klempí and Krupika, 2024). Those embeddings could still capture phonological and articulatory cues relevant to speech-disorder detection, so classification accuracy can be divergent from word-level transcription accuracy. This lack of correlation between the diagnostic capabilities of a model and its performance under transcription tasks highlights that while the classification performance of some models generalizes across languages, transcription is highly language-dependent.

In a preliminary analysis to evaluate model performance on tonal languages, we assessed disorder diagnosis capabilities on datasets of Taiwanese (Tsay, 2007) and Cantonese (Edwards and Beckman, 2008) speech. The results indicate a significant failure of current models to generalize to these linguistic contexts. As shown in Figure 8, performance is markedly degraded, with the best-performing model achieving a Micro F1 score below 0.10 in Taiwanese, while performance in Cantonese approached zero across all models.

Given that these evaluation datasets consisted entirely of speech from typically developing children, the low F1 scores reveal that the models systematically misdiagnosed the vast majority of samples as disordered. This suggests that the acoustic features these models rely on for disorder detection may be conflating the inherent tonal variations of these languages with pathological speech patterns. These findings underscore a critical limitation, indicating that current architectures are unable to accurately comprehend low-resource tonal languages within the SLP context.

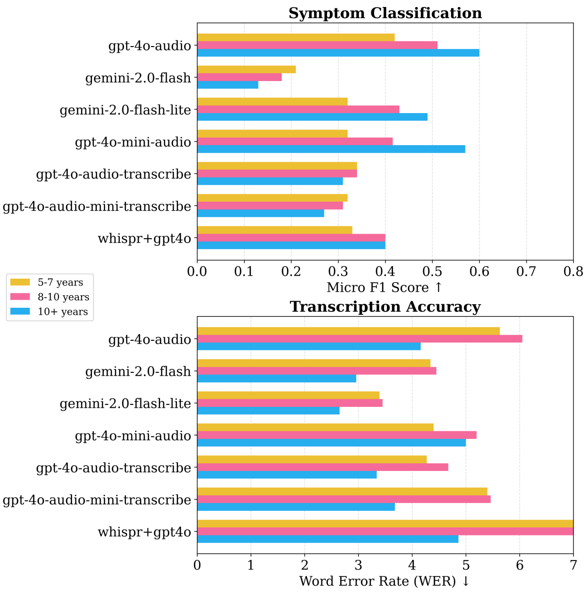

Figure 5: Model performance across age groups

The Effect of Age We analyzed model performance across three age ranges: early-elementary (5-7 years old), mid-elementary (8-10 years old), and post-elementary (10+ years old). The UtraSuite dataset provided structured age data for the children in its dataset and was used to run this analysis. The results reveal a distinct trend in symptom classification. Audio-native models demonstrate a significant drop in performance for younger children, as indicated by a lower Micro F1 Score for the 5-7 year-old group. In contrast, ASR+LLM pipelines maintain more consistent classification accuracy across all age ranges. For transcription accuracy, nearly all models performed best on the speech of the oldest children (10+). Performance gradually decreased for the younger age brackets.

The results showcase that variations in younger children's speech pose a significant challenge for audio-native models in complex classification tasks. The models are likely optimized for adult speech, and performance improves as children's speech patterns approach this norm. While transcription accuracy is also best for older children, the two-stage ASR+LLM approach appears more resilient to age-related speech variations for downstream analytical tasks like symptom classification, given that it relies on the model's text understanding capabilities to make its classification rather than audio understanding.

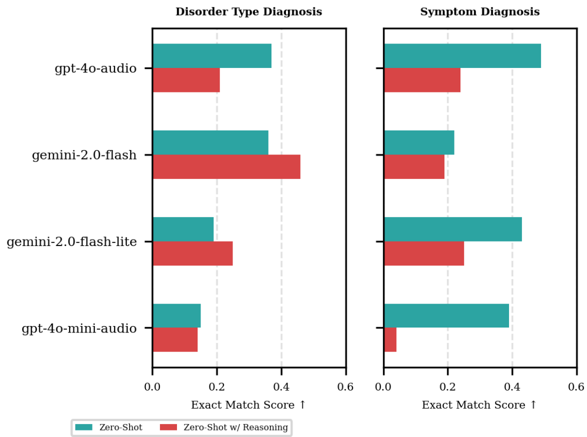

Figure 6: Compares model's classification performance with/without CoT prompting

The Effect of Reasoning We analysed the effects of introducing an explicit chain-of-thought (CoT) prompt across the Gemini 2.0 and GPT 4o model families. We saw that CoT systematically depressed F1 scores on the symptom diagnosis task and produced a mixed picture on the disorder type diagnosis, as can be seen in Figure 6. The pattern aligns with recent evidence that CoT can hamper tasks where the optimal decision boundary is compact or where answer formatting is unforgiving, because the additional reasoning tokens introduce distraction or bleed into the predicted label (Liu et al., 2024). However, even when CoT degrades final accuracy, the explicit rationale provides crucial insight into a model's failure modes. By inspecting the reasoning trace, we can diagnose why a model arrived at its conclusion, rather than simply observing the error itself. We analysed 200 reasoning traces, each for the 4 models for the Disorder-Type scenario. The CoTs are analysed along the following axes to reveal a clear profile of the model's reasoning process (Gandhi et al., 2025). A subset of the analysed traces can be found in Appendix G.

- **Subgoal Setting**, where a complex problem is broken down into manageable steps (e.g., "To solve this, we first need to...").
- **Rule following** or the application of taxonomical rules during classification (e.g. repeated errors in the last constonant indicate a phonological disorder rather than articulation).
- **Error Detection** or whether the model can recognize individual errors in speech, on which it will determine its classification.

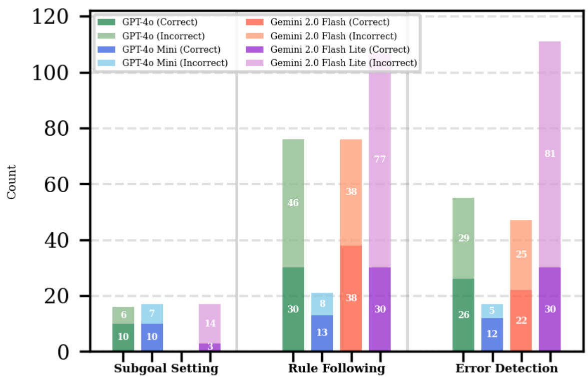

Figure 7: CoT Behaviour Analysis - Breakdown of behaviors showcased by the model. The bar for each model indicates the number of times the given behaviour is observed. Each bar is broken down by the times the given behaviour resulted in a correct/incorrect answer.

As shown in Figure 7, our analysis of CoT reasoning reveals key differences in the models' problem-solving approaches. Rule Following and Error Detection were the most prominent strategies employed across all models. However, the effective application of these strategies varied significantly. For GPT-4o, a majority of reasoning traces using Rule Following and Error Detection led to an incorrect final answer. This high rate of flawed application likely explains its reduced classification performance with CoT. Conversely, Gemini 2.0 Flash's improved performance correlates with more reliable reasoning; its application of Rule Following, for example, was correct in 50% of instances, compared to 39.5% for GPT-4o. This suggests that the fidelity of a model's reasoning directly impacts its accuracy. Interestingly, Gemini 2.0 Flash Lite exhibits the highest frequency of Rule Following and Error Detection behaviors but with the lowest accuracy (28% and 27%, respectively). Despite this, its classification performance improves, highlighting that the frequency of these model behaviours also contributes to overall model performance. This analysis indicates that for SLP tasks, CoT prompting improves performance only when the label space is limited with well-defined boundaries and the model can reliably apply rule-following and error-detection strategies in its reasoning traces.

Effect of Ensemble We implemented an ensemble strategy that combines predictions from multiple models. For classification tasks, the ensemble uses a majority voting system where the final prediction is determined by selecting the most frequently predicted label across all models, effectively leveraging the collective wisdom of multiple classifiers to reduce individual model biases and improve robustness. We created ensembles of three models to avoid tie scenarios where possible. If encountered, a tie is broken by arbitrarily selecting one of the predicted labels.

The ensemble analysis in Table 2 reveals a nuanced performance landscape. For the primary task of disorder diagnosis, a performance plateau is observed, with three distinct ensembles - the "Qwen Pure Ensemble" and both "Google + OpenAI" combinationsachieving similar F1 scores. However, significant differences emerge in the more granular classification tasks. The mixed-vendor Google + OpenAI ensembles demonstrate substantially better performance in symptom identification, achieving a Micro F1 score of 0.389, more than double that of the Qwen-based ensembles. Conversely, the Qwen ensembles show a slight advantage in disorder type classification. These findings suggest that a simple ensemble strategy does not guarantee superior performance, and that single model family ensembles are not guaranteed to be dominant. The effectiveness of an ensemble appears to be highly task-dependent and may benefit from combining models with diverse architectures, as seen with the Google and OpenAI ensembles' success in symptom classification. This highlights that the optimal ensemble configuration is not merely a combination of the strongest individual models but requires empirical testing to find complementary strengths for specific tasks.

The Effect of Background Noise We analyze model robustness by evaluating their performance with three added artificial perturbations, including road noise, classroom noise, and office noise. We aggregate the results to assess model performance. Each type of background noise was added at 20 dB. As shown in Figure 4, model performance degrades on average by 10 basis points across all models for disorder type diagnosis, while performance on symptom diagnosis remains virtually unchanged. These observations suggest that noise resilience is not strictly determined by model and inference architecture. It might instead be influenced by model-specific factors such as design, scale, and training data.

5 Conclusion & Future Work

Our empirical findings reveal that even the best-performing models fall short of clinical-grade reliability, revealing considerable room for improvement. This performance gap presents a significant clinical risk, as the deployment of an unreliable model could lead to patient misdiagnosis or delayed treatment. Our fine-tuning experiments with the Qwen2.5 family demonstrate that this performance can be substantially improved, highlighting the effectiveness of task-specific adaptation and the potential for developing specialized SLP models that generalize well across tasks.

Our robustness analysis identifies a consistent performance disparity favoring male speakers, contradicting the clinical principle of equitable care and highlighting the need for bias mitigation through strategies such as gender-balanced fine-tuning and targeted data augmentation. Cross-linguistic evaluations further demonstrate that audio-grounded models retain competitive diagnostic performance even when transcription quality degrades, suggesting that higher-order acoustic features support language-agnostic reasoning.

Future work could expand coverage to low-resource languages and neurodiverse populations, and evaluate model explanations for clinical faithfulness. We also plan to investigate privacy-preserving fine-tuning paradigms to facilitate deployment in sensitive pediatric settings. Collectively, these directions aim to bridge the gap between promising laboratory advances and the development of clinically robust, ethically sound AI systems for SLP.

Limitations and Ethical Considerations

Despite the comprehensive scope of our benchmark, several limitations and core ethical considerations warrant explicit discussion prior to any consideration of clinical deployment.

1. Privacy and Data Sensitivity

The ethical considerations surrounding privacy and consent are central to work in sensitive clinical domains. While this study was conducted on public datasets collected with informed consent and subsequently anonymized to protect participant privacy, our current setup does not incorporate a formal privacy-preserving learning or evaluation framework. Addressing this is essential to safeguard patient data and build trust among clinicians and families. As such, future work must investigate privacy-preserving fine-tuning paradigms to facilitate responsible deployment in pediatric settings.

2. Bias and Equity

A primary finding of this work is the identification of significant performance disparities, which present a direct challenge to the principle of equitable care.

- Gender Bias: Our evaluation revealed a consistent performance disparity that favors male speakers across multiple models. This indicates an urgent need for bias mitigation strategies, such as gender-balanced fine-tuning and targeted data augmentation, which were not implemented in this study but are critical next steps.

- Data Representation: The datasets employed are drawn primarily from English and French speakers, leading to an underrepresentation of other languages and dialects. This limits the generalizability of our findings and risks creating tools that are not effective for more linguistically and culturally diverse populations.

- Annotation Bias: The annotation process, though guided by an SLP professional, required prioritizing the most prominent disorder when multiple conditions co-occurred, introducing a degree of subjectivity.

3. Clinical Reliability and Validation

A core ethical requirement for deployment is robust clinical validation.

- Performance Gap: This study establishes a benchmark for clinical viability based on existing FDA-approved systems (F1 scores of

0.80 to 0.85). Our findings demonstrate that even the best-performing models currently fall short of this standard.

- Risk of Misdiagnosis: This performance gap represents a significant clinical risk, as the deployment of an unreliable model could lead to patient misdiagnosis or delayed treatment. The results underscore that current MLMs are not yet clinically robust and require substantial further development and validation before they can be considered for supportive roles in real-world SLP workflows.

Acknowledgments

SK acknowledges support by NSF 2046795 and 2205329, IES R305C240046, ARPA-H, the MacArthur Foundation, Schmidt Sciences, Stanford HAI, and Google. NH acknowledges the National AI Institute for Exceptional Education (Institute of Education Sciences, U.S. Department of Education, through Grant 22298673 (NSF)).

References

Michael D. Abràmoff, Patrick T. Lavin, Mary Birch, Nicholas Shah, and John C. Folk. 2018. Pivotal trial of an autonomous AI–based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1:39.

American Speech-Language-Hearing Association. 2016. Speech sound disorders: Articulation and phonology. `https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/`. [Practice Portal].

Nicholas R. Benway, Jonathan L. Preston, Elaine R. Hitchcock, Adam Salekin, Harsh Sharma, and Tara McAllister. 2022. PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of R. In *Proceedings of Interspeech 2022*, pages 2408–2412, Incheon, Republic of Korea. International Speech Communication Association (ISCA).

Anish Bhattacharya and 1 others. 2024. Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14(1):15573.

Centers for Disease Control and Prevention. 2024. Why act early if youre concerned about development? `https://www.cdc.gov/ncbddd/actearly/whyActEarly.html`. Accessed May 10, 2025.

Claire Cordella, Manuel J. Marte, Hantian Liu, and Swathi Kiran. 2025. An introduction to machine learning for speech-language pathologists: Concepts, terminology, and emerging applications. *Perspectives of the ASHA Special Interest Groups*, 10(2):432–450.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, and 1 others. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Jan Edwards and Mary E. Beckman. 2008. Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics*, 22(12):937–956.

Paul E Engelhardt and 1 others. 2021. A review of automated speech and language features for assessment of cognitive and thought disorders. *Frontiers in Psychology*.

Ahmed Eshky, Mário S. Ribeiro, Jane Cleland, Korin Richmond, Zoe Roxburgh, James Scobbie, and Alan Wrench. 2018. Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *Proceedings of Interspeech*, pages 2342–2346, Hyderabad, India. International Speech Communication Association (ISCA).

Zhaoxi Fan and 1 others. 2024. A benchmark for automatic speech recognition on child speech. In *Proc. INTERSPEECH*.

S. C. Fanni, A. Marcucci, F. Volpi, S. Valentino, E. Neri, and C. Romei. 2023. Artificial intelligence–based software with ce mark for chest x–ray interpretation: Opportunities and challenges. *Diagnostics*, 13(12):2020.

Adrian Florea, Xilin Jiang, Nima Mesgarani, and Xiaofan Jiang. 2025. Exploring finetuned audio-llm on heart murmur features. *Preprint*, arXiv:2501.13884.

M. E. Foster, A. L. Choo, and S. A. Smith. 2023. Speechlanguage disorder severity, academic success, and socioemotional functioning among multilingual and english children. *Frontiers in Psychology*, 14:1096145.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *Preprint*, arXiv:2503.01307.

D. Gibbard, L. Coglan, and J. MacDonald. 2004. Costeffectiveness analysis of a preschool speech and language therapy service. *International Journal of Language & Communication Disorders*, 39(1):1–11.

John Heilmann and 1 others. 2023. Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*.

E. R. Hitchcock, D. Harel, and T. M. Byun. 2015. Social, emotional, and academic impact of residual speech errors in school-age children. *Seminars in Speech and Language*, 36(4):283–294.

IBM Granite Team. 2025. Granite speech model documentation. `https://www.ibm.com/granite/docs/models/speech/`. Version last updated 14 May 2025; accessed 15 May 2025.

Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roee Aharoni, Bernd Bohnet, and Mor Geva. 2023. A comprehensive evaluation of tool-assisted generation strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13856–13878, Singapore. Association for Computational Linguistics.

Wataru Kawakami, Keita Suzuki, and Junichiro Iwasawa. 2025. Stabilizing reasoning in medical llms with continued pretraining and reasoning preference optimization. *Preprint*, arXiv:2504.18080.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. *Preprint*, arXiv:2302.03241.

Ondej Klempí and Radim Krupika. 2024. Analyzing wav2vec 1.0 embeddings for cross-database parkinsons disease detection and speech features extraction. *Sensors*, 24(17):5520. Shows wav2vec embeddings generalize across Italian and English PD datasets.

Jessica M. Lammert, Angela C. Roberts, Ken McRae, Laura J. Batterink, and Blake E. Butler. 2025. Early identification of language disorders using natural language processing and machine learning: Challenges and emerging approaches. *Journal of Speech, Language, and Hearing Research*, 68(2):705–718.

Marie-Thérèse Le Normand. 1997. Early morphological development in french children. In A. S. Olofsson and Sven Strömqvist, editors, *Cross-Linguistic Studies of Dyslexia and Early Language Development*, pages 59–79. Office for Official Publications of the European Communities, Luxembourg.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimers Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. In *Proc. INTERSPEECH*.

Umer Maqsood and 1 others. 2024. Large language models for dysfluency detection in stuttered speech. *arXiv preprint arXiv:2406.11025*.

Chirag Nagpal, Subhashini Venugopalan, Jimmy Tobin, Marilyn Ladewig, Katherine Heller, and Katrin Tomanek. 2025. Speech recognition with llms adapted to disordered speech using reinforcement learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *Preprint*, arXiv:2404.19733.

Leena G Pillai, Kavya Manohar, Basil K Raju, and Elizabeth Sherly. 2024. Multistage fine-tuning strategies for automatic speech recognition in low-resource languages. *Preprint*, arXiv:2411.04573.

Giacomo Sanguedolce and 1 others. 2024. Universal speech disorder recognition: Towards a foundation model for cross-pathology generalisation. `https://aphasia.talkbank.org/publications/2024/Sanguedolce24b.pdf`. Accessed May 2025.

Phyllis Schneider, Denyse Hayward, and Rebecca V. Dubé. 2006. Storytelling from pictures using the edmonton narrative norms instrument. *Journal of Speech-Language Pathology and Audiology*, 30:224–238.

Alyssa Schwartz and 1 others. 2024. Perceptions of artificial intelligence and chatgpt by speech-language pathologists and students. *American Journal of Speech-Language Pathology*.

Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. Personalizing asr for dysarthric and accented speech with limited data. In *Interspeech 2019*, pages 784–788.

Patti Simon and Sara Rosenbaum. 2016. *Speech and language disorders in children: Implications for the social security administration's supplemental security income program*. National Academies Press.

Hyewon Suh, Aayushi Dangol, Hedda Meadan, Carol A. Miller, and Julie A. Kientz. 2024. Opportunities and challenges for ai-based support for speech-language pathologists. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '24, New York, NY, USA. Association for Computing Machinery.

Jane Tsay. 2007. Construction and automatization of an Minnan child speech corpus with some research findings. *Computational Linguistics and Chinese Language Processing*, 12(4):411–442.

Dawn A. Tucker and Stella A. McKinnon. 2020. Addressing the shortage of speech–language pathologists in rural areas: Barriers and solutions. *Journal of Rural Health*, 36(4):620–628.

U.S. National Institute on Deafness and Other Communication Disorders. 2025. Quick statistics about voice, speech, language.

Yvette Wren, Linda L. Miller, Tim J. Peters, Ashley Emond, and Susan Roulstone. 2016. Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study. *J Speech Lang Hear Res*, 59(4):647–73.

Qimin Yang, Rongsheng Wang, Jiexin Chen, Runqi Su, and Tao Tan. 2024. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise. *Preprint*, arXiv:2407.11536.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset Statistic

Table 1: Dataset statistics

Dataset	# Children	# Samples	Age Range
Ultrasuite	66	8338	5–13
ENNI	377	16546	4–9
LeNormand (French)	17	329	3–8
PERCEPT-GFTA	350	3664	6–17

B Additional Related Works

AI in Speech Language Pathology Assessment The use of LLMs in clinical speech-language assessment has gained increasing attention in recent years. Several recent studies have demonstrated the utility of LLMs in detecting and characterizing speech and language disorders. For instance, Bhattacharya et al. (2024) showed that pre-trained LLMs could effectively identify both the presence and type of aphasia, suggesting that these models can serve as viable tools for clinical screening and diagnosis of language disorders. Beyond perception studies, a growing body of technical literature examines the use of speech and language features for automated assessment. Engelhardt et al. (2021) reviewed computational features used to assess cognitive and thought disorders, highlighting the relevance of acoustic and linguistic cues in differential diagnosis. Similarly, Heilmann et al. (2023) demonstrated that automatic language sample analysis tools can support clinical workflows, providing reliable linguistic metrics with reduced human effort.

LLMs for Disordered Speech Analysis A recent survey of SLPs and graduate students revealed a combination of cautious optimism and skepticism regarding the integration of LLMs such as ChatGPT into diagnostic and therapeutic workflows (Schwartz et al., 2024). These practitioner attitudes highlight critical socio-technical barriers to the clinical adoption of AI-driven systems in speech-language pathology. Recent research has explored the adaptation of LLMs for disordered speech processing.Sanguedolce et al. (2024) proposed a more generalized framework by fine-tuning Whisper on a dataset of stroke patients, resulting in a universal disordered-speech detection model. Their approach exhibited strong generalization across multiple neurological conditions, underscoring the potential of foundation models for broad-spectrum clinical speech applications.

C Additional Analysis

The Effect of Fewshot Examples The results of the GPT-4 family across the first three scenarios under few-shot prompting indicate that few-shot examples do not consistently enhance the model's intrinsic capabilities; the benefits of prompting are not uniformly evident. For instance, while few-shot prompting significantly improves the performance of GPT-4o-Mini-Transcribe and GPT-4o-Transcribe in Scenario 1, it leads to a reduction in accuracy in Scenario 2. Given that the examples provided to the model describe the expected transcription, few-shot prompting may bias the model to transcribe what the child was expected to say rather than accurately transcribe what the child is saying, which would explain the reduced diagnostic accuracy in Scenario 2. Our observations are consistent with prior findings on text-only LLMs (Jacovi et al., 2023).

D Fine-tuning details

D.1 Single Finetuning Round

We perform supervised fine-tuning on three models, including Qwen2-Audio 7B, Qwen2.5-Omni 3B, and Qwen2.5-Omni 7B using LLaMA-Factory framework (Zheng et al., 2024). We reuse the same UltraSuite dataset, but filtered out children whose speech in the evaluation set to avoid data contamination. We set the same fine-tuning hyperparameters across all models. Specifically, we used a LoRA rank of 32 with a LoRA alpha of 64, applying LoRA to all linear layers. The maximum token length was 4096, and training

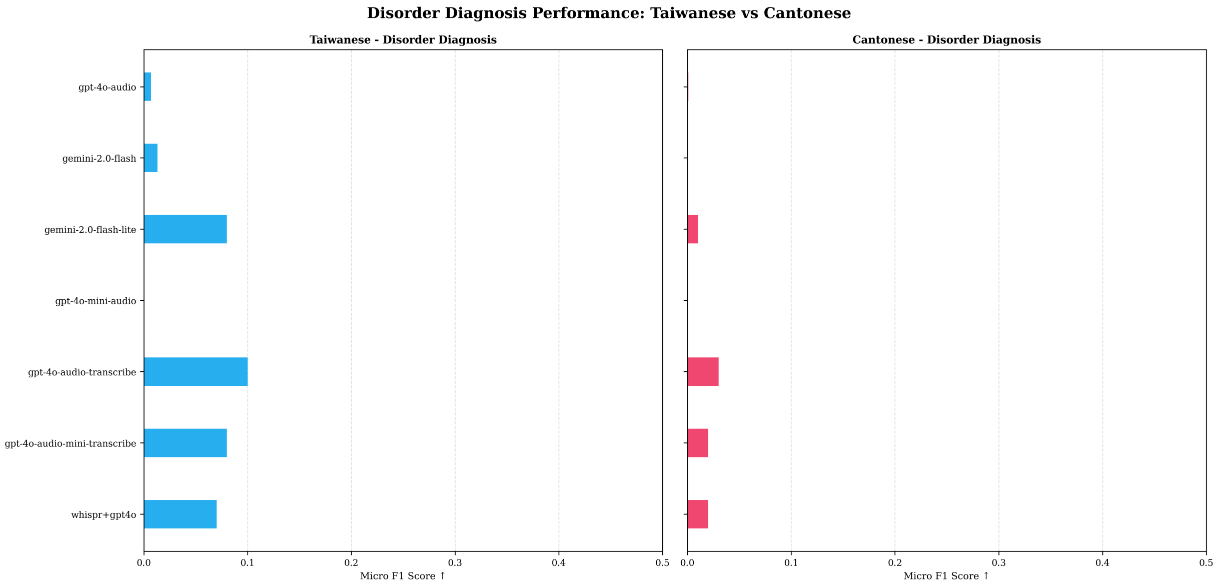

Figure 8: Disorder Diagnosis performance on tonal languages

was performed with a batch size of 32 for three epochs. We adopted a learning rate of 0.0001 with cosine scheduling and a warm-up ratio of 0.1. Regarding the prompts used for the three ablation settings for fine-tuning models, we present them as follows.

1. ASR-only without asterisk: `<audio>Transcribe this sound into text.`

2. ASR-only with asterisk: `<audio>Transcribe this sound into text. If the speech is disordered, please mark the words with an asterisk.`

D.2 Multiple Finetuning Rounds

To investigate the models capacity for self-evolution, we conducted an iterative finetuning experiment in which the training data were generated by the model itself. The procedure is summarized as follows. In each iteration, we prompted the current model to generate up to 2,000 samples comprising both normal and disordered textspeech pairs. The details of the data generation process are described below. The generated data were then used to finetune the same model, after which the updated model was evaluated using our proposed framework. This procedure was repeated for five iterations. We use the Qwen 2.5 Omni 3B model for this experiment, and the finetuning hyperparameters were kept identical to those used in the single-round finetuning experiment described above.

Data generation process. We employed the ViLLM framework to deploy the current model and prompt it to produce both disordered and normal text and corresponding speech. We use simple sentences from the Tatoeba dataset[2] as the input text. The prompt used to generate text and speech is as below.

- Normal text and speech: `Repeat exactly the following sentence and do not alter it. Return only the sentence and nothing else. The sentence is: {transcript}`

- Disordered text and speech: `Assuming you are a child with speech disordered. {disorder_description} Generate a corresponding transcript with the '{disorder_type} disorder' when the child say: "{transcript}". Return only the transcript text and nothing else.`

[2]https://huggingface.co/datasets/Helsinki-NLP/tatoeba_mt

If the model failed to generate disordered text, we manually introduced disorder markers into the text (including addition, substitution, omission, and stuttering). We then instructed the model to regenerate the disordered text and corresponding speech using the prompt as above. All generated audio was further processed through a *child-en* pipeline, which applied a series of signal processing operations (including pitch shifting, low-frequency reduction, high-frequency boosting, and slight speed-up) to approximate the characteristics of childrens speech. Subsequently, each sample was verified for detecting disorder using the same current model. If verification failed, generation was repeated up to five times; if it still failed, the sample was discarded, and the process moved on to the next sample. The prompt we used for verification is presented as follows.

```
You are a highly experienced Speech-Language Pathologist (SLP). An audio recording of a childs voice will be
    ↪  provided. The child is attempting to say: "\{transcript\}". Based on your professional expertise,
    ↪ analyze the childs speech production in the recording to determine whether it indicates a potential
    ↪ speech-language disorder (addition, substitution, omission, stuttering). Respond with one of the
    ↪ following labels only: `speech\_disorder`, `typically\_developing`. Do not include any explanation
    ↪ or additional text in your response.}
```

We present the results of the self-evolving experiment in Table 2. As shown, the transcription task exhibits a clear reduction in word error rate (WER), which can be attributed to the availability of the ground-truth transcript for each generated speech sample. In contrast, the performance on disorder diagnosis remains largely unchanged across iterations. This is expected, as the same model is used to verify whether the speech is disordered, thereby providing no additional learning signal. Consequently, the models performance on this task does not surpass that of the original model. A promising direction for future work is to employ a stronger verification model and/or incorporate supervision on the disorder type, which may yield richer signals for iterative finetuning.

E Detailed Results

2 presents our evaluation results. Values are reported as mean ś bootstrap standard deviation, providing both the point estimate and a measure of uncertainty in the metric estimates.

F Prompts

Scenario 1: Binary Classification

```
You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided,
    ↪ typically consisting of a speech prompt from a pathologist followed by a child's repetition. The
    ↪ prompt the child is trying to repeat is as follows: {words}. Based on your professional expertise:
    ↪ 1. Assess the child's speech in the recording for signs of typical development or potential speech-
    ↪ language disorder. 2. Conclude your analysis with one of the following labels only: typically
    ↪ developing or speech disorder. 3. Provide your response as a single letter without any additional
    ↪ explanation, commentary, or unnecessary text.
```

Scenario 2: ASRBased Classification

```
You are a highly experienced Speech-Language Pathologist (SLP). An audio recording is provided to you,
    ↪ typically consisting of a speech prompt from a pathologist followed by a child's repetition. Based
    ↪ on your expertise transcribe the child's speech into text. Do not make any assumptions about the
    ↪ words the child is expected to say. Only transcribe based on the words that the child actually says.
    ↪  Only respond with the text transcription, no other text or commentary.
```

Scenario 3: Transcription Accuracy

```
You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided,
    ↪ typically consisting of a speech prompt from a pathologist followed by a child's repetition. Based
    ↪ on your expertise transcribe the child's speech into text. Try to understand what the child is
    ↪ expected to say. And only respond with the transcription of the child's speech. Not the pathologist'
    ↪ s prompt or any other commentary. Only respond with the text transcription, no other text,
    ↪ commentary or punctuations.
```

Scenario 4: Disorder Type Classification

```
You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided,
    ↪ typically consisting of a speech prompt from a pathologist followed by a child's repetition. The
    ↪ prompt text the child is trying to repeat is as follows: {words}. Based on your professional
    ↪ expertise: 1. Assess the child's speech in the recording for signs of typical development or
    ↪ potential speech-language disorder. 2. Conclude your analysis with one of the following labels only:
    ↪  A - 'typically developing' (child's speech patterns and development are within normal age-
    ↪ appropriate ranges), B - 'articulation' (difficulty producing specific speech sounds correctly, such
    ↪  as substituting, omitting, or distorting sounds), C - 'phonological' (difficulty understanding and
    ↪ using the sound system of language, affecting sounds of a particular type). 3. Provide your response
    ↪  as a single letter without any additional explanation, commentary, or unnecessary text
```

Table 2: Detailed Model Performace Results

Model	Micro F1 ↑				Macro F1 ↑				Exact Match ↑						
	Disorder Diag.	ASR Diag.	Disorder Type	Symptom Diag.	Disorder Diag.	ASR Diag.	Disorder Type	Symptom Diag.	Disorder Diag.	ASR Diag.	Disorder Type	Symptom Diag.	WER ↓	MER ↓	WIP ↑
gemini-2.0-flash-lite	0.457 ±0.015	0.4590 ±0.011	0.180 ±0.012	0.415 ±0.016	0.341 ±0.010	0.3146 ±0.012	0.174 ±0.011	0.209 ±0.024	0.457 ±0.015	0.4590 ±0.011	0.180 ±0.012	0.415 ±0.016	0.51 ±0.016	0.51 ±0.016	0.49 ±0.016
gemini-2.0-flash	0.458 ±0.014	0.4600 ±0.013	0.294 ±0.015	0.180 ±0.013	0.330 ±0.009	0.3151 ±0.011	0.202 ±0.011	0.072 ±0.005	0.458 ±0.014	0.4600 ±0.013	0.294 ±0.015	0.180 ±0.013	1.00 ±0.000	1.00 ±0.000	0.00 ±0.000
gpt-4o-mini-audio	0.204 ±0.011	0.4838 ±0.009	0.125 ±0.011	0.433 ±0.017	0.208 ±0.009	0.3711 ±0.008	0.114 ±0.008	0.112 ±0.005	0.204 ±0.011	0.4838 ±0.009	0.125 ±0.011	0.433 ±0.017	5.64 ±0.066	0.96 ±0.006	0.04 ±0.006
gpt-4o-audio	0.439 ±0.014	0.4590 ±0.014	0.298 ±0.015	**0.507 ±0.016**	0.208 ±0.005	0.3163 ±0.015	0.166 ±0.007	0.189 ±0.009	0.439 ±0.014	0.4590 ±0.014	0.298 ±0.015	**0.507 ±0.016**	5.40 ±0.117	0.68 ±0.015	0.32 ±0.015
gpt-4o-mini-transcribe	0.456 ±0.018	0.4600 ±0.012	0.358 ±0.016	0.310 ±0.016	0.340 ±0.012	0.3168 ±0.013	0.274 ±0.015	0.209 ±0.014	0.456 ±0.018	0.4600 ±0.012	0.358 ±0.016	0.310 ±0.016	0.37 ±0.015	0.37 ±0.015	0.63 ±0.015
gpt-4o-transcribe	0.474 ±0.017	0.4600 ±0.011	0.382 ±0.016	0.325 ±0.015	0.373 ±0.014	0.3202 ±0.010	**0.320 ±0.017**	0.181 ±0.019	0.474 ±0.017	0.4600 ±0.011	0.382 ±0.016	0.325 ±0.015	1.00 ±0.000	1.00 ±0.000	0.00 ±0.000
whisper-gpt4o	0.473 ±0.015	0.4611 ±0.015	**0.399 ±0.016**	0.395 ±0.016	0.369 ±0.013	0.3190 ±0.014	0.245 ±0.013	**0.263 ±0.015**	0.473 ±0.015	0.4611 ±0.015	**0.399 ±0.016**	0.395 ±0.016	**0.33 ±0.018**	**0.29 ±0.015**	**0.71 ±0.015**
qwen2.5-omni-7b	0.455 ±0.006	0.713 ±0.008	0.345 ±0.005	0.163 ±0.005	**0.425 ±0.007**	0.4675 ±0.007	0.202 ±0.007	0.105 ±0.007	0.455 ±0.006	0.713 ±0.008	0.345 ±0.005	0.163 ±0.005	2.17 ±0.009	0.74 ±0.009	0.22 ±0.009
qwen2.5-omni-3b	**0.556 ±0.005**	0.738 ±0.007	0.396 ±0.016	0.155 ±0.016	0.401 ±0.008	0.4283 ±0.006	0.273 ±0.015	0.097 ±0.016	**0.556 ±0.005**	0.738 ±0.007	0.396 ±0.016	0.155 ±0.016	4.90 ±0.150	0.79 ±0.034	0.19 ±0.023
qwen2-audio-7b-instruct	0.447 ±0.005	0.743 ±0.006	0.335 ±0.005	0.097 ±0.003	0.206 ±0.002	0.4803 ±0.005	0.203 ±0.003	0.083 ±0.008	0.447 ±0.005	0.743 ±0.006	0.335 ±0.005	0.097 ±0.003	0.86 ±0.007	0.46 ±0.001	0.52 ±0.002
qwen-audio-chat	0.000 ±0.000	**0.747 ±0.004**	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.4276 ±0.008	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	**0.747 ±0.004**	0.000 ±0.000	0.000 ±0.000	1.87 ±0.068	0.58 ±0.002	0.40 ±0.002
phi-multimodal	0.325 ±0.005	0.724 ±0.009	0.371 ±0.005	0.127 ±0.004	0.255 ±0.004	**0.484 ±0.009**	0.181 ±0.002	0.086 ±0.003	0.325 ±0.005	0.724 ±0.009	0.371 ±0.005	0.127 ±0.004	2.54 ±0.169	0.37 ±0.004	0.51 ±0.004
granite-speech-3.3-8b	0.000 ±0.000	0.746 ±0.003	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.4528 ±0.004	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.746 ±0.003	0.000 ±0.000	0.000 ±0.000	9.11 ±0.117	0.58 ±0.005	0.41 ±0.005
granite-speech-3.3-2b	0.000 ±0.000	0.739 ±0.005	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.4564 ±0.003	0.000 ±0.000	0.000 ±0.000	0.000 ±0.000	0.739 ±0.005	0.000 ±0.000	0.000 ±0.000	2.34 ±0.018	0.71 ±0.002	0.25 ±0.002
granite-speech-3.2-8b	0.000 ±0.000	0.721 ±0.007	0.007 ±0.001	0.200 ±0.005	0.000 ±0.000	0.4797 ±0.006	0.007 ±0.001	0.096 ±0.002	0.000 ±0.000	0.721 ±0.007	0.007 ±0.001	0.200 ±0.005	2.25 ±0.107	0.46 ±0.004	0.48 ±0.003
Finetuned Models with Asterisk															
qwen2.5-omni-7b(ft)	**0.971±0.003**	0.631±0.020	**0.971±0.003**	**0.971±0.003**	**0.951±0.003**	0.511±0.010	**0.951±0.002**	**0.951±0.002**	**0.951±0.003**	0.631±0.020	**0.951±0.002**	**0.951±0.002**	1.40±0.011	0.52±0.006	0.41±0.009
qwen2.5-omni-3b(ft)	0.912±0.005	0.626±0.015	0.911±0.005	0.912±0.005	0.890±0.003	0.497±0.010	0.890±0.003	0.890±0.003	0.890±0.003	0.626±0.015	0.890±0.003	0.890±0.003	0.97±0.014	0.53±0.005	0.39±0.007
qwen2-audio-instruct(ft)	0.314±0.006	**0.689±0.018**	0.132±0.002	0.065±0.002	0.015±0.014	**0.596±0.005**	0.214±0.004	0.073±0.003	0.314±0.014	**0.689±0.018**	0.214±0.004	0.073±0.003	**0.58±0.017**	**0.43±0.004**	**0.50±0.008**
Finetuned Models without Asterisk															
qwen2.5-omni-7b(ft)	**0.67 ± 0.008**	**0.60 ± 0.012**	**0.40 ± 0.011**	**0.34 ± 0.007**	**0.81 ± 0.013**	**0.49 ± 0.015**	**0.28 ± 0.016**	**0.16 ± 0.016**	**0.67 ± 0.013**	**0.60 ± 0.012**	**0.40 ± 0.016**	**0.34 ± 0.016**	1.76 ± 0.019	0.46 ± 0.005	0.47 ± 0.008
qwen2.5-omni-3b(ft)	0.54 ± 0.009	0.59 ± 0.011	0.36 ± 0.010	0.16 ± 0.008	0.35 ± 0.013	0.46 ± 0.014	0.24 ± 0.016	0.08 ± 0.013	0.54 ± 0.013	0.59 ± 0.011	0.36 ± 0.016	0.16 ± 0.013	0.95 ± 0.014	0.49 ± 0.006	0.43 ± 0.007
qwen2-audio-instruct(ft)	0.36 ± 0.005	0.55 ± 0.004	0.27 ± 0.008	0.08 ± 0.007	0.01 ± 0.002	**0.49 ± 0.006**	0.05 ± 0.015	0.01 ± 0.009	0.36 ± 0.005	0.55 ± 0.004	0.27 ± 0.015	0.08 ± 0.009	**0.52 ± 0.015**	**0.38 ± 0.004**	**0.56 ± 0.006**
Fewshot Prompting															
gpt-4o-mini-audio	0.23	0.09	-	-	0.19	0.03	-	-	0.23	0.09	-	-	1.58	0.65	0.28
gpt-4o-audio	0.71	**0.12**	-	-	0.21	**0.04**	-	-	0.71	**0.12**	-	-	1.73	**0.62**	**0.30**
gpt-4o-mini-transcribe	**0.72**	0.01	-	-	0.44	0.01	-	-	**0.72**	0.01	-	-	1.08	0.80	0.12
gpt-4o-transcribe	**0.72**	0.01	-	-	**0.45**	0.01	-	-	**0.72**	0.01	-	-	**1.01**	0.79	0.14
whisper-gpt4o	**0.72**	0.01	-	-	0.24	0.00	-	-	**0.72**	0.01	-	-	1.89	0.77	0.16
Performance across multiple finetuning iterations															
Iteration 0 (baseline)	**0.555 ± 0.027**	0.738 ± 0.002	0.390 ± 0.029	0.150 ± 0.033	0.401 ± 0.021	0.428 ± 0.002	0.264 ± 0.019	0.094 ± 0.021	**0.555 ± 0.026**	0.738 ± 0.003	0.390 ± 0.016	0.150 ± 0.028	4.90 ± 0.15	**0.787 ± 0.034**	0.185 ± 0.023
Iteration 1	0.527 ± 0.032	0.740 ± 0.001	0.537 ± 0.031	0.123 ± 0.028	0.361 ± 0.020	0.425 ± 0.002	0.255 ± 0.020	0.106 ± 0.018	0.527 ± 0.017	0.740 ± 0.002	0.537 ± 0.034	0.123 ± 0.022	2.77 ± 0.23	0.865 ± 0.017	0.114 ± 0.025
Iteration 2	0.529 ± 0.034	0.740 ± 0.003	0.505 ± 0.022	0.140 ± 0.021	**0.407 ± 0.033**	0.425 ± 0.003	0.265 ± 0.024	0.122 ± 0.030	0.529 ± 0.032	0.740 ± 0.001	0.505 ± 0.032	0.140 ± 0.016	2.68 ± 0.11	0.851 ± 0.028	0.126 ± 0.030
Iteration 3	0.526 ± 0.024	0.740 ± 0.003	0.552 ± 0.018	0.125 ± 0.027	0.379 ± 0.026	0.425 ± 0.002	0.239 ± 0.028	0.080 ± 0.030	0.526 ± 0.025	0.740 ± 0.002	0.552 ± 0.033	0.125 ± 0.029	2.63 ± 0.17	0.854 ± 0.029	0.121 ± 0.017
Iteration 4	0.528 ± 0.018	0.740 ± 0.003	0.553 ± 0.029	0.128 ± 0.024	0.388 ± 0.028	0.425 ± 0.002	0.254 ± 0.021	0.088 ± 0.019	0.528 ± 0.034	0.740 ± 0.001	0.553 ± 0.027	0.128 ± 0.021	2.42 ± 0.22	0.852 ± 0.034	0.122 ± 0.023
Iteration 5	0.531 ± 0.026	0.740 ± 0.003	0.554 ± 0.017	0.140 ± 0.024	0.385 ± 0.033	0.425 ± 0.003	0.250 ± 0.018	0.097 ± 0.029	0.531 ± 0.019	0.740 ± 0.002	0.554 ± 0.027	0.140 ± 0.029	**2.14 ± 0.14**	0.846 ± 0.023	0.129 ± 0.023
Model Ensemble Performance															
Google + OpenAI Ensemble G. Gemini 2.0 Flash, G. Gemini 2.0 Flash Lite, OpenAI GPT-4o	0.455	-	0.313	0.389	0.319	-	0.215	0.184	0.455	-	0.313	0.389	0.96	0.57	0.03
Google + OpenAI Variant G. Gemini 2.0 Flash, OpenAI GPT-4o, OpenAI GPT-4o Mini	0.455	-	0.313	0.389	0.319	-	0.215	0.184	0.455	-	0.313	0.389	0.92	0.54	0.07
Qwen Pure Ensemble Qwen2.5 Omni 3B, Qwen2 Audio 7B Instruct, Qwen2.5 Omni 7B	0.455	-	0.339	0.163	0.426	-	0.200	0.09	0.455	-	0.339	0.163	5.16	3.31	0.13
Qwen + OpenAI Hybrid Qwen2.5 Omni 7B, Qwen2 Audio 7B Instruct, OAI GPT 4o	0.403	-	0.338	0.167	0.300	-	0.189	0.109	0.403	-	0.338	0.167	1.88	1.75	0.3
Performance by Age															
Gemini 2.0 Flash															
pre-elementary	0.7629 ±0.0291	0.7700 ±0.0164	0.475 ±0.037	0.1329 ±0.012	0.4326 ±0.0094	0.4350 ±0.0347	0.299 ±0.026	0.0557 ±0.008	0.7629 ±0.0291	0.7700 ±0.0281	0.475 ±0.037	0.1329 ±0.012	4.346 ±0.115	1.000 ±0.000	0.000 ±0.000
elementary	0.423 ±0.023	0.4500 ±0.0309	-	0.1836 ±0.014	0.303 ±0.013	0.3103 ±0.0184	-	0.0729 ±0.009	0.423 ±0.023	0.4500 ±0.0325	-	0.1836 ±0.014	4.452 ±0.129	0.998 ±0.001	0.000 ±0.000
post-elementary	0.280 ±0.026	0.2800 ±0.0281	0.200 ±0.024	0.1329 ±0.011	0.228 ±0.019	0.2188 ±0.0190	0.156 ±0.019	0.0557 ±0.007	0.280 ±0.026	0.2800 ±0.0229	0.200 ±0.024	0.1329 ±0.011	**2.960 ±0.088**	0.998 ±0.001	0.000 ±0.000
Gemini 2.0 Flash Lite															
pre-elementary	0.7288 ±0.0302	0.7700 ±0.0167	0.306 ±0.028	0.4930 ±0.018	0.4214 ±0.0101	0.4350 ±0.0345	0.245 ±0.021	**0.2102 ±0.013**	0.7288 ±0.0302	0.7700 ±0.0201	0.306 ±0.028	0.4930 ±0.018	**3.290 ±0.091**	1.000 ±0.000	0.000 ±0.000
elementary	0.428 ±0.023	0.4450 ±0.0256	-	0.4361 ±0.019	0.322 ±0.015	0.3080 ±0.0427	-	**0.2419 ±0.015**	0.428 ±0.023	0.4450 ±0.0188	-	**0.4361 ±0.019**	**3.458 ±0.102**	0.990 ±0.002	0.000 ±0.000
post-elementary	0.271 ±0.026	0.2800 ±0.0194	0.106 ±0.017	0.4930 ±0.017	0.222 ±0.018	0.2188 ±0.0118	0.119 ±0.017	**0.2102 ±0.012**	0.271 ±0.026	0.2800 ±0.0288	0.106 ±0.017	**0.4930 ±0.017**	2.650 ±0.076	0.992 ±0.002	0.000 ±0.000
GPT-4o Audio															
pre-elementary	0.7490 ±0.0294	0.7700 ±0.0154	0.561 ±0.034	**0.6049 ±0.021**	0.4281 ±0.0096	0.4350 ±0.0347	0.278 ±0.021	0.1648 ±0.011	0.7490 ±0.0294	0.7700 ±0.0243	0.561 ±0.034	**0.6049 ±0.021**	5.630 ±0.155	1.000 ±0.000	0.000 ±0.000
elementary	0.414 ±0.023	0.4600 ±0.0223	-	**0.5115 ±0.020**	0.296 ±0.012	0.3304 ±0.0308	-	0.2014 ±0.013	0.414 ±0.023	0.4600 ±0.0342	-	0.5115 ±0.020	6.052 ±0.181	0.996 ±0.001	0.000 ±0.000
post-elementary	0.271 ±0.026	0.2800 ±0.0260	0.261 ±0.025	**0.6049 ±0.019**	0.219 ±0.018	0.2188 ±0.0110	0.166 ±0.016	0.1648 ±0.010	0.271 ±0.026	0.2800 ±0.0229	0.261 ±0.025	0.6049 ±0.019	4.162 ±0.112	1.000 ±0.000	0.000 ±0.000
GPT-4o Mini Audio															
pre-elementary	0.0984 ±0.0202	0.7900 ±0.0317	0.232 ±0.028	0.5734 ±0.020	0.0886 ±0.0180	0.5200 ±0.0477	0.140 ±0.013	0.1226 ±0.010	0.0984 ±0.0202	0.7900 ±0.0199	0.232 ±0.028	0.5734 ±0.020	4.416 ±0.130	1.000 ±0.000	0.000 ±0.000
elementary	0.148 ±0.017	0.4700 ±0.0204	-	0.4164 ±0.018	0.137 ±0.015	0.3629 ±0.0201	-	0.1161 ±0.011	0.148 ±0.017	0.4700 ±0.0293	-	0.4164 ±0.018	5.166 ±0.162	1.000 ±0.000	0.000 ±0.000
post-elementary	0.083 ±0.016	0.3200 ±0.0320	0.072 ±0.015	0.5734 ±0.018	0.071 ±0.014	0.2784 ±0.0392	0.078 ±0.016	0.1226 ±0.009	0.083 ±0.016	0.3200 ±0.0161	0.072 ±0.015	0.5734 ±0.018	5.008 ±0.145	1.000 ±0.000	0.000 ±0.000
GPT-4o Mini Transcribe															
pre-elementary	0.7540 ±0.0294	0.7700 ±0.0306	0.491 ±0.038	0.2797 ±0.016	0.4297 ±0.0096	0.4555 ±0.0373	0.333 ±0.034	0.1310 ±0.011	0.7540 ±0.0294	0.7700 ±0.0182	0.491 ±0.038	0.2797 ±0.016	5.404 ±0.151	1.000 ±0.000	0.000 ±0.000
elementary	0.422 ±0.022	0.4500 ±0.0251	-	0.3180 ±0.017	0.316 ±0.016	0.3180 ±0.0469	-	0.2609 ±0.016	0.422 ±0.022	0.4500 ±0.0299	-	0.3180 ±0.017	5.466 ±0.168	1.000 ±0.000	0.000 ±0.000
post-elementary	0.304 ±0.027	0.2800 ±0.0326	0.254 ±0.026	0.2797 ±0.015	0.269 ±0.023	0.2188 ±0.0323	0.179 ±0.019	0.1310 ±0.010	0.304 ±0.027	0.2800 ±0.0210	0.254 ±0.026	0.2797 ±0.015	3.684 ±0.101	1.000 ±0.000	0.000 ±0.000
GPT-4o Transcribe															
pre-elementary	0.7670 ±0.0292	0.7700 ±0.0324	0.496 ±0.037	0.3112 ±0.017	0.4842 ±0.0319	0.4555 ±0.0244	0.347 ±0.037	0.1432 ±0.012	0.7670 ±0.0292	0.7700 ±0.0172	0.496 ±0.037	0.3112 ±0.017	4.276 ±0.122	1.000 ±0.000	0.000 ±0.000
elementary	0.457 ±0.025	0.4550 ±0.0193	-	0.3475 ±0.018	0.369 ±0.019	0.3204 ±0.0368	-	0.2256 ±0.014	0.457 ±0.025	0.4550 ±0.0304	-	0.3475 ±0.018	4.674 ±0.140	1.000 ±0.000	0.000 ±0.000
post-elementary	0.295 ±0.024	0.2800 ±0.0264	0.289 ±0.029	0.3112 ±0.016	0.258 ±0.020	0.2188 ±0.0204	0.204 ±0.020	0.1432 ±0.011	0.295 ±0.024	0.2800 ±0.0338	0.289 ±0.029	0.3112 ±0.016	3.342 ±0.096	1.000 ±0.000	0.000 ±0.000
Whisper + GPT-4o															
pre-elementary	0.7161 ±0.0304	0.7750 ±0.0346	0.522 ±0.032	0.4091 ±0.019	0.3269 ±0.0673	0.4575 ±0.0275	0.398 ±0.037	0.2184 ±0.014	0.7161 ±0.0304	0.7750 ±0.0281	0.522 ±0.032	0.4091 ±0.019	10.898 ±0.254	0.980 ±0.003	0.000 ±0.000
elementary	0.454 ±0.024	0.4500 ±0.0212	-	0.4000 ±0.019	0.384 ±0.021	0.3103 ±0.0112	-	**0.2785 ±0.017**	0.454 ±0.024	0.4500 ±0.0322	-	0.4000 ±0.019	10.226 ±0.280	**0.976 ±0.004**	0.000 ±0.000
post-elementary	0.316 ±0.028	0.2800 ±0.0230	0.393 ±0.029	0.4091 ±0.017	0.288 ±0.025	0.2188 ±0.0451	0.275 ±0.019	0.2184 ±0.012	0.316 ±0.028	0.2800 ±0.0152	0.393 ±0.029	0.4091 ±0.017	4.866 ±0.135	**0.944 ±0.005**	0.000 ±0.000
Effect of Gender															
Gemini 2.0 Flash															
Female	0.2986 ±0.015	0.2600 ±0.018	0.2380 ±0.014	0.1120 ±0.010	0.2453 ±0.013	0.2063 ±0.014	0.1779 ±0.011	0.0469 ±0.006	0.2986 ±0.015	0.2600 ±0.014	0.2380 ±0.014	0.1120 ±0.010	0.956 ±0.021	0.703 ±0.018	0.236 ±0.011
Male	0.5640 ±0.019	0.5600 ±0.015	0.4360 ±0.018	0.2000 ±0.014	0.3648 ±0.016	0.3590 ±0.012	0.2511 ±0.014	0.0750 ±0.008	0.5640 ±0.019	0.5600 ±0.011	0.4360 ±0.018	0.2000 ±0.014	0.967 ±0.022	0.697 ±0.018	0.234 ±0.011
Gemini 2.0 Flash Lite															
Female	0.3096 ±0.016	0.2600 ±0.013	0.0900 ±0.009	0.4880 ±0.018	0.2617 ±0.014	0.2063 ±0.018	0.0996 ±0.009	0.2149 ±0.014	0.3096 ±0.016	0.2600 ±0.011	0.0900 ±0.009	0.4880 ±0.018	0.843 ±0.019	0.712 ±0.018	0.223 ±0.010
Male	0.5600 ±0.018	0.5600 ±0.016	0.2380 ±0.015	0.3640 ±0.017	0.2503 ±0.013	0.3590 ±0.019	0.1978 ±0.012	0.1780 ±0.013	0.5600 ±0.018	0.5600 ±0.010	0.2380 ±0.015	0.3640 ±0.017	0.846 ±0.019	0.706 ±0.018	0.223 ±0.010
GPT-4o Audio															
Female	0.2959 ±0.015	0.2650 ±0.015	0.2360 ±0.014	**0.6400 ±0.020**	0.1654 ±0.011	0.2139 ±0.019	0.1154 ±0.010	0.2217 ±0.015	0.2959 ±0.015	0.2650 ±0.012	0.2360 ±0.014	**0.6400 ±0.020**	2.348 ±0.051	0.691 ±0.017	0.251 ±0.012
Male	0.5580 ±0.019	0.5550 ±0.010	0.4140 ±0.018	**0.4440 ±0.019**	0.2438 ±0.012	0.3569 ±0.012	0.2037 ±0.013	0.1761 ±0.012	0.5580 ±0.019	0.5550 ±0.012	0.4140 ±0.018	**0.4440 ±0.019**	2.214 ±0.048	0.675 ±0.017	0.263 ±0.012
GPT-4o Mini Audio															
Female	0.1616 ±0.012	0.2850 ±0.011	0.1080 ±0.010	0.5200 ±0.019	0.1898 ±0.012	0.2469 ±0.014	0.1150 ±0.010	0.1286 ±0.011	0.1616 ±0.012	0.2850 ±0.013	0.1080 ±0.010	0.5200 ±0.019	3.017 ±0.062	0.729 ±0.019	0.213 ±0.010
Male	0.2180 ±0.014	0.5900 ±0.011	0.1720 ±0.013	0.3680 ±0.017	0.2051 ±0.011	0.4298 ±0.014	0.1247 ±0.010	0.0981 ±0.009	0.2180 ±0.014	0.5900 ±0.020	0.1720 ±0.013	0.3680 ±0.017	2.739 ±0.055	0.728 ±0.019	0.211 ±0.010
GPT-4o Mini Transcribe															
Female	0.2904 ±0.015	0.2650 ±0.013	0.3420 ±0.016	0.2960 ±0.015	0.2377 ±0.013	0.2139 ±0.019	0.2608 ±0.014	0.1793 ±0.013	0.2904 ±0.015	0.2650 ±0.020	0.3420 ±0.016	0.2960 ±0.015	1.716 ±0.041	0.766 ±0.020	0.182 ±0.009
Male	0.5660 ±0.019	0.5650 ±0.012	0.5000 ±0.019	0.2800 ±0.016	0.3855 ±0.017	0.3714 ±0.017	0.3432 ±0.016	0.2164 ±0.014	0.5660 ±0.019	0.5650 ±0.018	0.5000 ±0.019	0.2800 ±0.016	1.695 ±0.040	0.761 ±0.020	0.185 ±0.009
GPT-4o Transcribe															
Female	0.3370 ±0.016	0.2650 ±0.010	0.3420 ±0.016	0.3400 ±0.016	0.3061 ±0.015	0.2139 ±0.016	0.2609 ±0.014	0.1622 ±0.012	0.3370 ±0.016	0.2650 ±0.017	0.3420 ±0.016	0.3400 ±0.016	1.550 ±0.038	0.740 ±0.019	0.205 ±0.010
Male	**0.5760 ±0.020**	0.5600 ±0.013	0.4980 ±0.019	0.2480 ±0.015	**0.4086 ±0.018**	0.3590 ±0.019	0.3804 ±0.017	0.1700 ±0.012	**0.5760 ±0.020**	0.5600 ±0.017	0.4980 ±0.019	0.2480 ±0.015	1.525 ±0.037	0.735 ±0.019	0.210 ±0.010
Whisper + GPT-4o															
Female	**0.3507 ±0.017**	0.2650 ±0.010	**0.3780 ±0.017**	0.4320 ±0.018	**0.3259 ±0.016**	0.2139 ±0.012	**0.2670 ±0.015**	**0.2421 ±0.015**	**0.3507 ±0.017**	0.2650 ±0.017	**0.3780 ±0.017**	0.4320 ±0.018	**0.455 ±0.015**	**0.410 ±0.013**	**0.580 ±0.014**
Male	0.5120 ±0.019	0.5600 ±0.018	**0.5120 ±0.020**	0.3600 ±0.017	0.3956 ±0.017	0.3590 ±0.019	**0.3956 ±0.018**	**0.2848 ±0.016**	0.5120 ±0.019	0.5600 ±0.018	**0.5120 ±0.020**	0.3600 ±0.017	**0.421 ±0.014**	**0.395 ±0.012**	**605 ±0.015**
Performance Across Languages															
Gemini 2.0 Flash															
Cantonese	0.0000 ±0.000	0.0000 ±0.000	-	-	0.0000 ±0.000	0.0000 ±0.000	-	-	0.0000 ±0.000	0.0000 ±0.000	-	-	1.420 ±0.050	1.000 ±0.000	0.000 ±0.000
Dutch	**1.0000 ±0.000**	0.2300 ±0.014	-	-	0.3843 ±0.018	0.1870 ±0.011	-	-	**1.0000 ±0.000**	0.2300 ±0.014	-	-	4.708 ±0.135	1.000 ±0.000	0.000 ±0.000
French	**0.9970 ±0.002**	0.8400 ±0.021	-	-	0.5365 ±0.021	0.4565 ±0.019	-	-	0.2736 ±0.016	0.8400 ±0.021	-	-	1.196 ±0.041	0.809 ±0.011	0.135 ±0.010
Spanish	0.2540 ±0.015	0.0000 ±0.000	-	-	0.2043 ±0.014	0.0000 ±0.000	-	-	0.2540 ±0.015	0.0000 ±0.000	-	-	**1.542 ±0.051**	1.000 ±0.000	0.000 ±0.000
Taiwanese	0.0140 ±0.004	0.0000 ±0.000	-	-	0.0138 ±0.004	0.0000 ±0.000	-	-	0.0140 ±0.004	0.0000 ±0.000	-	-	1.420 ±0.050	1.000 ±0.000	0.000 ±0.000
Gemini 2.0 Flash Lite															
Cantonese	0.0120 ±0.003	0.0000 ±0.000	-	-	0.0119 ±0.003	0.0000 ±0.000	-	-	0.0120 ±0.003	0.0000 ±0.000	-	-	1.370 ±0.048	1.000 ±0.000	0.000 ±0.000
Dutch	0.9860 ±0.004	0.2300 ±0.013	-	-	**0.4965 ±0.020**	0.1870 ±0.012	-	-	0.9860 ±0.004	0.2300 ±0.013	-	-	3.276 ±0.101	0.998 ±0.001	0.000 ±0.000
French	0.9574 ±0.007	0.8400 ±0.022	-	-	0.3498 ±0.018	0.4565 ±0.018	-	-	0.1216 ±0.011	0.8400 ±0.022	-	-	0.877 ±0.032	0.752 ±0.014	0.176 ±0.012
Spanish	**0.3320 ±0.017**	0.0000 ±0.000	-	-	**0.3320 ±0.017**	0.0000 ±0.000	-	-	**0.3320 ±0.017**	0.0000 ±0.000	-	-	1.720 ±0.058	1.000 ±0.000	0.000 ±0.000
Taiwanese	0.0960 ±0.009	0.0000 ±0.000	-	-	0.0876 ±0.009	0.0000 ±0.000	-	-	0.0960 ±0.009	0.0000 ±0.000	-	-	1.370 ±0.048	1.000 ±0.000	0.000 ±0.000
GPT-4o Audio															
Cantonese	0.0020 ±0.001	0.0000 ±0.000	-	-	0.0013 ±0.001	0.0000 ±0.000	-	-	0.0020 ±0.001	0.0000 ±0.000	-	-	1.430 ±0.050	1.000 ±0.000	0.000 ±0.000
Dutch	0.9406 ±0.008	0.2300 ±0.012	-	-	0.3231 ±0.017	0.1870 ±0.014	-	-	0.9406 ±0.008	0.2300 ±0.012	-	-	5.663 ±0.165	1.000 ±0.000	0.000 ±0.000
French	0.9392 ±0.009	0.8350 ±0.023	-	-	0.5885 ±0.022	0.4550 ±0.021	-	-	0.1094 ±0.010	0.8350 ±0.023	-	-	1.234 ±0.043	0.849 ±0.012	0.105 ±0.009
Spanish	0.1940 ±0.013	0.0000 ±0.000	-	-	0.0952 ±0.010	0.0000 ±0.000	-	-	0.1940 ±0.013	0.0000 ±0.000	-	-	2.194 ±0.071	1.000 ±0.000	0.000 ±0.000
Taiwanese	0.0120 ±0.003	0.0000 ±0.000	-	-	0.0079 ±0.003	0.0000 ±0.000	-	-	0.0120 ±0.003	0.0000 ±0.000	-	-	1.430 ±0.050	1.000 ±0.000	0.000 ±0.000
GPT-4o Mini Audio															
Cantonese	0.0000 ±0.000	0.0000 ±0.000	-	-	0.0000 ±0.000	0.0000 ±0.000	-	-	0.0000 ±0.000	0.0000 ±0.000	-	-	2.150 ±0.075	1.000 ±0.000	0.000 ±0.000
Dutch	0.0332 ±0.006	0.2300 ±0.015	-	-	0.0214 ±0.005	0.1870 ±0.013	-	-	0.0332 ±0.006	0.2300 ±0.015	-	-	5.942 ±0.178	1.000 ±0.000	0.000 ±0.000
French	0.9574 ±0.007	0.8400 ±0.020	-	-	0.6215 ±0.023	0.4565 ±0.022	-	-	**0.7204 ±0.025**	0.8400 ±0.020	-	-	1.211 ±0.042	**0.861 ±0.011**	0.086 ±0.008
Spanish	0.0000 ±0.000	0.0000 ±0.000	-	-	0.0000 ±0.000	0.0000 ±0.000	-	-	0.0000 ±0.000	0.0000 ±0.000	-	-	5.330 ±0.159	1.000 ±0.000	0.000 ±0.000
Taiwanese	0.0640 ±0.008	0.0000 ±0.000	-	-	0.0401 ±0.006	0.0000 ±0.000	-	-	0.0640 ±0.008	0.0000 ±0.000	-	-	2.150 ±0.075	1.000 ±0.000	0.000 ±0.000
GPT-4o Mini Transcribe															
Cantonese	0.0220 ±0.005	0.0000 ±0.000	-	-	0.0215 ±0.005	0.0000 ±0.000	-	-	0.0220 ±0.005	0.0000 ±0.000	-	-	1.060 ±0.037	1.000 ±0.000	0.000 ±0.000
Dutch	0.9545 ±0.007	0.2300 ±0.016	-	-	0.4884 ±0.020	0.1870 ±0.015	-	-	0.9545 ±0.007	0.2300 ±0.016	-	-	**2.678 ±0.088**	1.000 ±0.000	0.000 ±0.000
French	0.8967 ±0.011	0.8400 ±0.023	-	-	0.7349 ±0.026	0.4565 ±0.017	-	-	0.1702 ±0.013	0.8400 ±0.023	-	-	0.871 ±0.031	0.762 ±0.013	0.160 ±0.011
Spanish	0.2540 ±0.015	0.0000 ±0.000	-	-	0.2296 ±0.014	0.0000 ±0.000	-	-	0.2540 ±0.015	0.0000 ±0.000	-	-	1.662 ±0.055	1.000 ±0.000	0.000 ±0.000
Taiwanese	0.0960 ±0.009	0.0000 ±0.000	-	-	0.0876 ±0.009	0.0000 ±0.000	-	-	0.0960 ±0.009	0.0000 ±0.000	-	-	1.060 ±0.037	1.000 ±0.000	0.000 ±0.000
GPT-4o Transcribe															
Cantonese	**0.0320 ±0.006**	0.0000 ±0.000	-	-	**0.0310 ±0.006**	0.0000 ±0.000	-	-	**0.0320 ±0.006**	0.0000 ±0.000	-	-	**1.020 ±0.036**	1.000 ±0.000	0.000 ±0.000
Dutch	0.9441 ±0.008	0.2300 ±0.011	-	-	0.4856 ±0.020	0.1870 ±0.016	-	-	0.9441 ±0.008	0.2300 ±0.011	-	-	2.734 ±0.091	1.000 ±0.000	0.000 ±0.000
French	0.9058 ±0.010	0.8250 ±0.025	-	-	**0.7667 ±0.027**	0.4521 ±0.023	-	-	0.2340 ±0.015	0.8250 ±0.025	-	-	**0.807 ±0.029**	0.758 ±0.014	0.173 ±0.012
Spanish	0.2840 ±0.016	0.0000 ±0.000	-	-	0.2721 ±0.016	0.0000 ±0.000	-	-	0.2840 ±0.016	0.0000 ±0.000	-	-	1.664 ±0.056	1.000 ±0.000	0.000 ±0.000
Taiwanese	**0.1100 ±0.010**	0.0000 ±0.000	-	-	**0.0991 ±0.010**	0.0000 ±0.000	-	-	**0.1100 ±0.010**	0.0000 ±0.000	-	-	**1.020 ±0.036**	1.000 ±0.000	0.000 ±0.000
Whisper + GPT-4o															
Cantonese	0.0280 ±0.005	0.0000 ±0.000	-	-	0.0272 ±0.005	0.0000 ±0.000	-	-	0.0280 ±0.005	0.0000 ±0.000	-	-	1.740 ±0.061	1.000 ±0.000	0.000 ±0.000
Dutch	0.9161 ±0.010	0.2300 ±0.017	-	-	0.1594 ±0.013	0.1870 ±0.018	-	-	0.9161 ±0.010	0.2300 ±0.017	-	-	7.166 ±0.203	**0.939 ±0.006**	0.000 ±0.000
French	0.9088 ±0.010	0.8400 ±0.024	-	-	0.5184 ±0.020	0.4565 ±0.024	-	-	0.2310 ±0.015	0.8400 ±0.024	-	-	1.200 ±0.041	0.710 ±0.015	**0.190 ±0.013**
Spanish	0.2960 ±0.016	0.0000 ±0.000	-	-	0.0448 ±0.007	0.0000 ±0.000	-	-	0.2960 ±0.016	0.0000 ±0.000	-	-	2.314 ±0.075	**0.868 ±0.009**	0.000 ±0.000
Taiwanese	0.0860 ±0.009	0.0000 ±0.000	-	-	0.0792 ±0.008	0.0000 ±0.000	-	-	0.0860 ±0.009	0.0000 ±0.000	-	-	1.360 ±0.048	**0.988 ±0.007**	**0.011 ±0.002**
Performance Under Noisy Conditions															
gemini-2.0-flash	0.4720 ±0.018	0.4550 ±0.018	0.3320 ±0.016	0.1840 ±0.013	0.3207 ±0.015	0.3127 ±0.015	**0.2443 ±0.014**	0.0653 ±0.008	0.4720 ±0.018	0.4550 ±0.018	0.3320 ±0.016	0.1840 ±0.013	1.027 ±0.010	0.723 ±0.010	0.207 ±0.010
gemini-2.0-flash-lite	0.4720 ±0.018	0.4500 ±0.017	0.1480 ±0.012	0.4540 ±0.018	0.3271 ±0.016	0.3103 ±0.016	0.1438 ±0.011	0.2335 ±0.015	0.4720 ±0.018	0.4500 ±0.017	0.1480 ±0.012	0.4540 ±0.018	**0.849 ±0.010**	0.747 ±0.010	0.190 ±0.010
gpt-4o-audio	0.4700 ±0.018	0.4550 ±0.019	0.2900 ±0.015	**0.5120 ±0.019**	0.2188 ±0.013	0.3127 ±0.014	0.1557 ±0.012	0.1804 ±0.013	0.4700 ±0.018	0.4550 ±0.019	0.2900 ±0.015	**0.5120 ±0.019**	1.779 ±0.010	**0.710 ±0.010**	**0.222 ±0.010**
gpt-4o-mini-audio	0.0740 ±0.009	0.4700 ±0.020	0.0520 ±0.007	0.4240 ±0.018	0.0896 ±0.009	0.3428 ±0.017	0.0558 ±0.007	0.1012 ±0.010	0.0740 ±0.009	0.4700 ±0.020	0.0520 ±0.007	0.4240 ±0.018	2.863 ±0.010	0.763 ±0.010	0.177 ±0.010
gpt-4o-mini-transcribe	0.4880 ±0.019	0.4550 ±0.018	0.3420 ±0.016	0.3180 ±0.016	0.3594 ±0.017	0.3127 ±0.015	0.2101 ±0.013	0.1425 ±0.012	0.4880 ±0.019	0.4550 ±0.018	0.3420 ±0.016	0.3180 ±0.016	1.134 ±0.010	0.861 ±0.010	0.077 ±0.010
gpt-4o-transcribe	0.4980 ±0.019	0.4950 ±0.021	0.3660 ±0.017	0.3080 ±0.016	**0.3761 ±0.017**	0.3311 ±0.016	0.2274 ±0.014	0.1338 ±0.011	0.4980 ±0.019	0.4950 ±0.021	0.3660 ±0.017	0.3080 ±0.016	1.081 ±0.010	0.872 ±0.010	0.078 ±0.010
whisper-gpt4o	**0.5060 ±0.019**	0.4550 ±0.019	**0.4820 ±0.019**	0.4600 ±0.018	0.2695 ±0.014	0.3127 ±0.015	0.2338 ±0.014	**0.2476 ±0.015**	**0.5060 ±0.019**	0.4550 ±0.019	**0.4820 ±0.019**	0.4600 ±0.018	1.841 ±0.010	0.836 ±0.010	0.107 ±0.010

Scenario 5: SymptomClassification

```
You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided,
    ↪ typically consisting of a speech prompt from a pathologist followed by a child's repetition. The
    ↪ target phrase the child is attempting to repeat is: \{words\}. Based on your professional expertise,
    ↪  assess the child's speech in the recording and identify any abnormal features. These features can
    ↪ be one of the following: A - 'substitution' (the child replaces one word, syllable, or sound with
    ↪ another), B - 'omission' (the child omits a word, syllable, or sound), C - 'addition' (the child
    ↪ adds an extra word, syllable, or sound), D - 'typically developing' (the child's speech is
    ↪ appropriate for their age), or E - 'stuttering' (the child exhibits repetition, prolongation, or
    ↪ difficulty initiating speech). Provide your response as a single letter (A-E) only, without any
    ↪ additional explanation or commentary.
```

G Gemini 2.0 Flash Reasoning Traces

Below, we provide some examples of reasoning traces output by the Gemini 2.0 Flash model when tasked with disorder type diagnosis.

Identify the Target Sounds: The target words are "gorge," "gate," "get," and "luggage." The key sounds within these words include /g/, /r/, /e/, /l/.
Listen for Substitutions/Distortions: The child says "gore" for "gorge", "gape" for "gate", "get" appears correct, and "nugget" for "luggage".
Analyze the Error Patterns: "Gore" for "gorge" seems to be a mispronunciation or substitution of the vowel and omitting the final sound /dʒ/. "Gape" for "gate" also a mispronunciation or substitution of the vowel sound /e/. "Nugget" for "luggage" involves multiple sound changes. The initial $/l/$ is replaced with $/n/$, the vowel /ʌ/ is present, and the final sound /I/ is replaced by /t/.
Determine the Type of Speech Error: The sound changes observed are primarily substitutions and omissions, indicating difficulties with producing specific sounds accurately.
Classify the Speech Pattern: Based on the analysis, the most appropriate classification is articulation, as the errors relate to the physical production of individual speech sounds.

Okay, let's analyze the child's speech based on the prompt "owrow awraw ooroo". The prompt presents a challenge, focusing on vowel sounds within a simple consonant-vowel-consonant (CVC) structure. The child's repetitions are "O wow", "A wa", and "Uu Uu".
* **"owrow" becomes "O wow"**: This indicates a possible difficulty in producing the /r/ sound in the middle of the word. The vowel sounds are relatively close but the /r/ is replaced by a /w/. * **"awraw" becomes "A wa"**: Similar to the first attempt, the child simplifies or substitutes the /r/ sound with a /w/. The vowel sounds are again reasonably close. * **"ooroo" becomes "Uu Uu"**: This repetition is more successful, with the child accurately producing the vowel sound, but there is no attempt to use the /r/. This suggests that the child is either not hearing the /r/ sound, or they are hearing it but unable to produce it.
The consistent difficulty with the /r/ sound suggests an articulation issue specifically related to that sound. The vowel sounds seem relatively well approximated, pointing away from significant phonological problems where a whole sound class (e.g., all back vowels) would be affected.
Conclusion: B - 'articulation'

Okay, let's analyze the child's speech based on the recording.
The prompt was "lie reek hair strewn."
The child said: "lie reek hair shm."
* **lie:** Correct production. * **reek:** Correct production. * **hair:** Correct production. * **strewn:** "shm". This indicates a cluster reduction, where the /str/ consonant cluster is reduced to /shm/.
The error on "strewn" demonstrates difficulty with consonant clusters, specifically /str/. There's no evidence of distortions, substitutions, or omissions on single phonemes in the other words. The error is on a later-developing sound.
Given that the child is having difficulty with consonant clusters it suggests that a phonological process is at play.
Conclusion:
C - 'phonological'

Okay, let's analyze this speech sample.
The child attempts the phrase "coot packer toot tie".
* **"coot"**: The child says "koot" which sounds accurate. * **"packer"**: The child says "packa". The "r" sound is missing or distorted. * **"toot"**: The child says "toot" which sounds accurate. * **"tie"**: The child says "tie" which sounds accurate.
The primary error observed is with the "r" sound in "packer." It's possible that the "r" is simply omitted, or it may be a distorted production. The other sounds are articulated relatively clearly. This difficulty with the "r" sound suggests an issue with producing a specific sound, rather than a broader issue with the phonological system.
Conclusion: B - 'articulation'

Okay, let's analyze the child's speech.
The prompt is: "You get milk from cows." The child's response is: "You get milk from dows."
The child substitutes the /k/ sound in "cows" with a /d/ sound, resulting in "dows." This is a substitution error affecting a specific sound. There's no evidence of errors impacting an entire sound class (e.g., all fricatives, all velar sounds). The error involves difficulty with the articulation of the /k/ sound. Therefore, the most appropriate label is 'articulation'.
Reasoning: * The child's speech contains a sound substitution. * The substitution affects one specific sound: /k/ becoming /d/. * The errors involve specific sounds, not an entire sound class (which would indicate a phonological issue).
Answer: B

Okay, let's analyze the child's speech based on the provided audio.
The prompt was: "kai meek poke cow" The child responded with: "tai, meet, poat, dow"
Here's a breakdown:
* **kai -> tai:** The child substitutes /k/ with /t/. * **meek -> meet:** The vowel sounds are similar; no concerning error. * **poke -> poat:** The child substitutes /k/ with /t/. * **cow -> dow:** The child substitutes /k/ with /d/.
The child consistently substitutes /k/ with /t/ and /k/ with /d/. This pattern suggests a difficulty with the /k/ sound. Since the errors are consistent across multiple instances of the /k/ sound, it points to a potential issue with the phonological system rather than just a simple articulation error.
Conclusion:
C - 'phonological'

The child's speech contains several sound errors. "Crab" is produced as "gap," suggesting difficulty with the /kr/ consonant cluster. "Biscuits" is produced as "bisits," which contains a sound distortion of /k/. "Helicopter" is produced as "hedoctor," involving multiple sound substitutions and omissions. These errors point to difficulties with the articulation and phonological representation of sounds.
ANSWER: B

Okay, let's analyze the child's repetition of "DDK ppp ttt kkk ptk patticake."
The prompt is designed to assess the child's ability to produce rapid, alternating sounds (diadochokinesis or DDK) and to articulate specific phonemes in isolation and in a word.
* **DDK (ppp ttt kkk):** The child attempts the DDK task and does show signs of struggling. The child seems to understand the prompt, and can repeat the p, t and k sounds. However the sounds appear jumbled together. * **ptk:** The child attempts ptk. The child appears to jumble this prompt. * **patticake:** The child repeats this prompt several times. The child shows great difficulty producing this word and does not enunciate the word well.
Based on these observations, the child exhibits difficulties with articulation and sequencing sounds. While single sounds can be produced, transitioning between them appears challenging. This suggests the primary difficulty is in the motor production of the sounds.
Conclusion: B - 'articulation'

Okay, I've listened to the audio. Here's my analysis:
Reasoning:
The target words contain the // phoneme (as in "sh"). The child says "shore," then what sounds like "tosh," then the child attempts to repeat "puss." I hear what sounds like a substitution of // with /s/ in the word "puss," making it sound similar to the target word of "puss." This consistent substitution of // with /s/ (or perhaps a distortion close to /s/) across multiple words containing the // phoneme suggests a potential phonological pattern, specifically affecting fricative sounds.
Conclusion:
C - 'phonological'

Okay, let's analyze the child's speech based on the recording. - **Sock:** The child's repetition of "sock" sounds accurate. - **Shoop:** The child's repetition of "shoop" sounds accurate. - **Shire:** The child's repetition of "shire" sounds like "show." This suggests a potential difficulty with the /r/ sound after the /sh/ sound. - **Shallow:** The child's repetition of "shallow" sounds like "show." This suggests a potential difficulty with the /l/ sound after the /sh/ sound.
Based on these observations, the child demonstrates difficulty with the production of the /r/ and /l/ sounds in the context of the /sh/ blend. There is evidence of a potential speech sound disorder. Given the errors are consistent with particular sound classes, it is most indicative of a phonological pattern.
Final Answer: The final answer is C

Okay, let's analyze the child's speech. The prompt "owrow awraw ooroo" contains several instances of the "r" sound in different syllable positions. The SLP in the recording notes that there was barely an "r" in the first attempt and it sounded like "all to ohs." The subsequent attempt on "owrow" was considered much better. The final attempt on "ooroo" appears to be a distortion, possibly an "r" sound replaced with a vowel or gliding sound. Based on the limited data, there is an apparent difficulty with the production of the /r/ sound. This could indicate an articulation issue, where the child struggles with the motor movements required for the /r/ sound. Since the difficulty is centered on one specific sound, articulation is the most appropriate label.
Answer: B

Okay, I will analyze the provided audio and provide a diagnosis based on my expertise as an SLP.
Here's my analysis:
The child's repetition of "A dozen green ghosts" sounds generally clear. However, a closer listening reveals a potential difficulty with the 'th' sound, as in "ghosts". The repetition of "ghosts" is not as clear compared to the phrase, "dozen green".
Reasoning: 1. The child produces "dozen green" relatively accurately. 2. There is a possible sound substitution affecting the /s/ sound. 3. Due to the specific nature of the sound, and a mild misarticulation it suggests a more localized articulation difficulty rather than a broader phonological issue.
Therefore, my diagnosis is:
B - 'articulation'

Okay, let's analyze the child's speech based on the provided audio.
The target words are "book boy". The child says something that sounds like "buh... book... boy."
Here's a breakdown of what I'm hearing: * The child seems to produce the "b" sound correctly at the beginning of both words. * For "book," they seem to be attempting the vowel sound and the "k" sound. It doesn't sound perfect, but it could be due to typical developmental imprecision, or it could indicate a slight articulation issue with the "k" sound. * For "boy," the "b" sounds correct, and the vowel sound sounds within an acceptable range.
Given the information and the presence of the "b" at the beginning of both words, the rest of the sounds aren't clear enough to indicate a Phonological delay. However, there are some slight errors in production.
Therefore, my assessment is:
ANSWER: B - 'articulation'

Okay, I will listen to the audio and analyze the child's speech based on the prompt "sheen shake sham shock."
Analysis:
The child appears to have difficulty with the initial // sound in all four words. While the vowel sounds seem relatively intact, the consistency of the // distortion suggests a pattern rather than an isolated error. It is difficult to determine the exact nature of the distortion without a more thorough assessment (e.g., is it a substitution, omission, or distortion?). The consistency across multiple words with the same sound indicates a possible phonological issue.
Conclusion:
C - 'phonological'

Okay, let's analyze the child's speech sample.
The prompt words are: "die," "got," "hag," and "Hague".
Here's a breakdown of potential issues:
* **"die"**: The child repeats "die" which sounds accurate. * **"got"**: The child repeats "got" which sounds accurate. * **"hag"**: The child repeats "hag" which sounds accurate. * **"Hague"**: The child repeats "Hake" which involves changing the final sound.
The one slight discrepancy could be related to the place of articulation. Because this appears to be a single instance, I will not classify this as a phonological issue. I'm confident that this instance points to a lack of production of the "g" sound in "Hague."
Conclusion:
B - 'articulation'

H Computational Resources

To ensure reproducibility and transparency, we detail the computational resources required for running the SLPHelm framework as follows. Closed-source models were accessed via API, with an average inference time of approximately 30 minutes per model. In contrast, open-source models required significantly more time, averaging 3–4 hours per model on high-performance hardware. All open-source experiments were conducted using an NVIDIA H200 GPU with 140GB of VRAM, enabling efficient handling of large-scale model inference and evaluation.

I AI Tool Usage Declaration

We used ChatGPT to format paragraphs and correct grammatical mistakes. We are responsible for the content and quality of the study.